Универзитет у Београду
Математички факултет

Проблем недостајућих података: утицај на статистичко закључивање

Мастер рад

Данијел Алексић
1009/2021

Ментор:

проф. др Бојана Милошевић, ванредни професор

Чланови комисије:

др Марко Обрадовић, доцент (председник комисије)

др Марија Цупарић, доцент

Датум одбране:

22.09.2022.

Садржај

Предговор

Овај рад за циљ има да изложи свеобухватан преглед приступа за руковање недостајућим подацима, који су свакодневна појава у пракси, као и да истражи утицај различитих импутационих метода на један од тестова вишедимензионе нормалности. Организован је у пет глава.

Глава 1 служи упознавању са основним концептима анализе недостајућих података. Изложена је конвенција о ознакама које ће бити коришћене, а након тога је изложен задатак статистике и начелни поступак статистичког закључивања. Затим је описано како присуство недостајућих података ствара проблем при статистичком закључивању и дати су адекватни примери. Након тога је изложена формална класификација образаца и механизама недостајања, што је илустровано на адекватном примеру. Дат је пример теста који тестира претпоставку потпуно случајног недостајања, а затим је понуђена груба класификација метода за решавање проблема недостајућих података.

У глави 2 изложени су основни видови статистичког закључивања, прво на основу комплетних, а затим и података који имају недостајуће вредности. Уведен је појам *игнорабилности* механизма недостајања за сваки од та четири начина. Свима њима је својствено то да се претпоставља да су подаци узорковани из извесне познате расподеле вероватноћа, те се на ту расподелу ослања у свим тим типовима закључивања.

Глава 3 нуди други угао гледања на проблем, који често заобилази потребу за спецификацијом модела над подацима: *импутацију недостајућих података*. Импутација представља процес попуњавања недостајућих поља заменским вредностима. Објашњено је које су предности и мане у односу на задавање модела над подацима, те је дат кратак преглед фундаменталних резултата у области импутације. Извршена је класификација на експлицитно и имплицитно моделовање, те је за сваку од класа дато по неколико примера. Уочени су проблеми који се тичу неадекватних стандардних грешака и других оцена параметара добијених на основу импутираних података, а који су последица неурачунавања неодређености коју имамо о вредности коју импутирамо.

Решење за овај проблем јесте тзв. *вишеструка импутација*, којој је посвећена глава 4. За почетак се даје преглед неких начина за урачунавање неодређености о недостајућој вредности у импутациону процедуру, а затим се даје кратак историјски преглед метода. Поставља се општи бајесовски контекст разматрања и дискутују се могућа поједностављења. Након тога, објашњава се начин симулације апостериорног очекивања и дисперзије, као и услови под којима су те симулације валидне. Уводе се Рубинова правила. Затим се посматра фреквенционистичка природа бајесовских оцена, и то при случајном узорку или случајном недостајању. Изводе се расподеле оцена у том контексту и дискутује се валидност такве (фреквенционистичке) процедуре. Након тога прелази се на интервалне оцене параметара, као и p-вредности тестова, и излажу се познати резултати у том правцу. Дискутују се слабе тачке и потенцијални (неопходни) нови правци истраживања. На крају главе дају се примери неких често коришћених метода и алгоритама којима се врши вишеструка импутација.

Пета глава посвећена је изучавању утицаја различитих импутационих метода на тестове вишедимензионе нормалности. Конкретно, посматрано је како се мења моћ BHEP теста уколико се он спроводи користећи допуњене податке, а коришћен је метод калибрације који се огледа у томе да се тестира не жељеним, већ оним нивоом значајности који даје очекивану меру теста. Спроведена је опсежна симулациона студија и дискутовани су резултати.

Изјава захвалности

Иако је испред мене и даље дугачак пут напредовања као математичара, сматрам да је стићи до ове тачке на којој се тренутно налазим ипак велики успех, те имам потребу захвалити се свима који су мом успеху допринели.

Пре свега, захваљујем се мојој прерано преминулој мајци Милофинки, која ме је још са три године научила да читам и пишем и која ми је до последњег дана била велика подршка. Сигуран сам да би се овом мом резултату радовала више него ја. Затим се захваљујем свом оцу Гојку, који ми представља сигуран ослонац и заштиту.

Хвала мојој менторки, проф. др Бојани Милошевић, којој се увек могу обратити са било каквим питањем, немајући страх од тога које је доба дана или ноћи. Њена помоћ била ми је од пресудног значаја при писању овог рада. Сви резултати за које се у раду каже да су нови, добијени су у сарадњи с њом. Хвала и свим осталим професорима са Катедре за вероватноћу и статистику Математичког факултета Универзитета у Београду, од којих сам много научио. Године проведене студирајући и радећи на тој катедри засигурно су период мог живота у којем сам највише напредовао. Посебно се захваљујем члановима комисије на корисним сугестијама.

Такође, захвалност дугујем и мојим колегама са Катедре за математику Факултета организационих наука Универзитета у Београду, који су били препуни разумевања за моје факултетске обавезе, због којих је моје тамошње ангажовање понекад морало и да претрпи.

На крају, хвала свим мојим пријатељима, који су више веровали у мене него ја сам.

Надам се да ће овај мој рад некоме послужити било као користан приручник за рад са недостајућим подацима, или као одскочна даска у његовом даљем научном напредовању. Мени је његово стварање засигурно помогло.

У Београду,
Септембра 2022. године,
Данијел Алексић.

Глава 1

Увод у проблем недостајућих података

1.1 О појмовима и ознакама

Још од почетака људске цивилизације људи се занимају за прикупљање, пренос и обраду података. Ипак, под речју *подаци* подразумевао се много широк скуп појмова. То су могле бити речи, песме, слике, цртежи, рукописи - просто било шта што се може пренети од човека до човека на било који начин.

У данашњој науци дошло је до благе формализације тог појма, али не сасвим. У машинском учењу, рецимо, под подацима се може подразумевати и слика, и звучни сигнал, као и много тога другог. Ипак, у статистици смо доста формалнији. Када кажемо да су нам дати подаци, подразумевамо да су нам дати *табеларни подаци*, што значи да нам је дата матрица $\mathbf{Y} = [y_{ij}]$ димензија $n \times K$, где је n број врста у матрици $\mathbf{Y}$, а K број њених колона.

Свака од врста матрице $\mathbf{Y}$ је облика $(y_{i1}, \ldots, y_{iK})$ и она заправо представља реализацију једног случајног вектора $(Y_1, \ldots, Y_K)$. Суштински, врсте у нашим подацима представљају реализације независних и једнако расподељених случајних величина и углавном ћемо их звати *опсервацијама* или *инстанцама*. Сваку од врста замишљамо као јединку на којој смо измерили вредности одговарајућих обележја, која код нас представљају колоне матрице $\mathbf{Y}$ и зовемо их *предикторима* или атрибутима. Неки од атрибута може бити категорички (нпр. пол), али ћемо такве на одговарајући начин нумерички кодирати.

Често ћемо наизменично користити ознаке $Y_1, \ldots, Y_K$ и за случајне величине и за реализоване колоне, али ће из контекста увек бити јасно на шта се мисли. За било коју од координата неке опсервације $(y_{i1}, \ldots, y_{iK})$ често ћемо користити израз *ћелија*, или *поље*, што никако не треба мешати са алгебарском структуром поља.

У неким ситуацијама користићемо мало другачије ознаке. Рецимо, када се бавимо линеарном регресијом, или, општије, било којим дискриминативним моделом, тада обично једну колону у подацима означимо са Y и зовемо *циљна променљива*, јер правимо модел који њу процењује на основу осталих. Тада је пракса да се за остале колоне користе ознаке $X_1, X_2, X_3, \ldots$. Промене ознака ћемо објашњавати како будемо упадали у такве ситуације, а трудићемо се да их буде што мање.

1.2 Задатак статистике

Као и за било коју другу математичку и уопште научну дисциплину, веома је тешко и незахвално прецизно дефинисати шта је њен задатак. То се увек, некако, зна. Ипак, угрубо говорећи, можемо рећи да је посао једног статистичара *да проналази правилности у подацима* и *да из њих изводи закључке*. Он то ради градећи разне *моделе*, који треба да представљају што бољу апроксимацију стварних односа међу подацима.

Једном статистичару у ту сврху доступни су многи статистички алати. Користи се велики број регресионих модела попут линеарне регресије, као и њених регуларизованих варијан-

ти попут назубљене или *lasso* регресије, метода потпорних вектора итд. Користе се класификациони алгоритми, разни методи за редукцију димензионалности, затим кластеризацију података и слично. Иза свега овога у позадини стоје фундаментални резултати теорије вероватноћа и математичке статистике, попут закона великих бројева, централне граничне теореме, Гливленко-Кантелијеве теореме, а оне се пак наслањају на теорију мере, функционалну анализу, и тако даље уназад до математичке логике.

Након што је направљен статистички модел, треба испитати његов квалитет. За то се користе разне мере квалитета модела, које у зависности од његове природе могу попримити разне облике. У честој су употреби средњеквадратна грешка, унакрсна ентропија, тачност, прецизност и одзив класификације итд. Дајмо пример.

Пример 1.1 (линеарна регресија)**.** Претпоставимо да имамо податке који су дати у матрици

$$[Y, \mathbf{X}] = \begin{bmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

и да на основу датих података желимо да направимо модел који описује зависност променљиве Y од променљивих[1] $\mathbf{X} = (X_1, \ldots, X_p)$. Како ми нисмо заинтересовани за моделовање заједничке расподеле Y и $\mathbf{X}$, већ само везе којом прва од друге зависи, то ћемо ми подразумевати да су нам X_j-ови константе и моделоваћемо *условну расподелу* $Y \mid \mathbf{X} = \mathbf{x}$. Код модела линеарне регресије претпоставља се модел облика

$$Y|\mathbf{X} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

где смо претпоставили да је $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, а да смо матрицу $\mathbf{X}$ допунили водећом колоном јединица. У расписаном облику, за свако i имамо да је

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

где претпостављамо да су X_{ij}-ови константе, јер моделујемо условну расподелу, а да су ε_i-јеви случајне величине које представљају грешке и задовољавају Гаус-Марковљеве услове центрираности, некорелисаности и хомоскедастичности. Често се претпоставља и нормална расподељеност грешака.

Под претпоставком да желимо да минимизујемо средњеквадратну грешку, може се показати да су оптималне оцене коефицијената дате изразом

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

Ова оцена је, дакле, оптимална[2] под претпоставком да је линеарни модел адекватан за податке, да желимо да минимизујемо средњеквадратну грешку и, посебно, да грешке задовољавају услове Гаус-Маркова. Међутим, реални подаци које смо ми добили не морају испуњавати те услове, те ће линеарнорегресиони модел бити веома неадекватан. То треба проверити.

Рецимо да желимо да проверимо некорелисаност грешака. Пракса је показала да је честа алтернатива томе да грешке задовољавају Гаус-Марковљеве услове то да оне прате ауторегресиону зависност:

$$\varepsilon_i = a\varepsilon_{i-1} + u_i,$$

где је константа, а низ (u_i) задовољава услове Гаус-Маркова. Тада се провера некорелисаности грешака своди на тестирање хипотезе

$$H_0 : a = 0,$$

за шта се може користити Дарбин-Вотсонова статистика

[1] Овде већ упадамо у први проблем са нотацијом: са $\mathbf{X}$ смо означили и случајни вектор и његову n-тоструку независну реализацију записану матрично. Из контекста је, надамо се, јасно када се мисли на шта од то двоје.

[2] Има најмању дисперзију од свих непристрасних оцена.

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

која при нултој хипотези има вредност 2. Наравно, e_i су резидуали модела: $e_i = Y_i - \hat{Y}_i$.

У претходном примеру илустровали смо шта значи формирати један статистички модел, као и како тестирати испуњеност једне од њихових претпоставки. Слично се може тестирати хомоскедастичност, по потреби нормална расподељеност итд. Није тешко генерализовати овакво закључивање и уочити да ће се при прављењу било каквог модела појавити потреба за формирањем неке тест статистике, констатовањем њене (асимптотске) расподеле при нултој хипотези и коришћењем тога за извођење закључака. Међутим, све ове процедуре почивају на томе да су сви подаци доступни, да се на основу њих може формирати било каква оцена или друга статистика, критичне области, интервали поверења итд.

Шта радити уколико подаци нису комплетни, то јест ако недостаје једна или више ћелија у подацима? На ово и на слична питања покушава да одговори грана статистике која се бави радом са недостајућим подацима.

1.3 Проблем недостајућих података

Дакле, нека су наши подаци дати матрицом Y коју смо увели на почетку ове главе, али нека неки од њих нису доступни за опажање. Веома је битно нагласити да **подаци који недостају постоје, имају своју вредност, само ми не можемо да их опазимо**. При покушају да срачунамо било шта, па рецимо оцене линеарнорегресионих коефицијената, резидуале или Дарбин-Вотсонову статистику, наићи ћемо на проблем да је то немогуће, јер просто не знамо шта треба да заменимо на неким местима у формули.

Ad hoc решење[3] јесте да се из података избаци свака врста којој недостаје барем једна ћелија. Тада кажемо да вршимо анализу само комплетних врста, што је у литератури на енглеском језику познато као *complete-case analysis*.

Ово може да представља огроман проблем. Уколико је некомплетних врста мало и уколико подаци недостају на приближно случајан начин (што ћемо касније детаљније деифинисати), бићемо на малом губитку. Знамо да дисперзија већине оцена опада са квадратом растојања, па ћемо просто, смањивши обим узорка, малкице повећати дисперзије наших оцена. Али, шта ако подаци недостају на начин који није случајан?

Пример 1.2. Једна од честих ситуација јесу тзв. цензурисани подаци. Рецимо да имамо вагу која је у стању да измери 100 килограма или мање. Уколико на њу стане особа која има већу масу, вага ће приказати 100kg и упозорење да је гранична маса прекорачена, па нећемо имати праву вредност масе за посматрану опсервацију. Ово је пример делимичног недостајања. Једноставности ради, претпоставићемо да ако је особа тежа од 100 килограма, вага приказује грешку, а не број 100. Јасно је да ово недостајање и није баш случајно.

Уколико из података који недостају на овакав или сличне начине избацимо све некомплетне врсте, упадамо у проблем јер смо изгубили репрезентативност узорка. Просто, у свету постоји много људи тежих од 100 килограма, а у нашем прерађеном узорку их неће бити. Ово ће да поквари оцене било каквих коефицијената у моделу и, још битније, *покварiће расподеле тест статистика*, па ће резултати многих тестирања бити погрешни. Рецимо, закључићемо да је неки коефицијент уз предиктор нула, а он није нула (Валдов тест), или ћемо закључити да су грешке некорелисане иако нису.

Штавише, поред проблема са контаминацијом оцена и „кварењем" расподела тест статистика, јасно је да губимо информацију. Можда наши подаци имају 20 предиктора, и у свакој некомплетној врсти недостаје само по једна ћелија. Тада, избацивањем сваке врсте губимо 19 познатих бројева, а суштински не знамо само 1. Све претходно речено наводи нас на то да брисање некомплетних врста и није баш најбоље решење за руковање недостајућим подацима.

[3] Које се често и користи, мада углавном не би требало.

Алтернатива томе било би нешто што је у литератури познато као *pairwise analysis*, а ми ћемо га звати приступ пар-по-пар[4]. Оваква процедура састоји се у томе да избацујемо некомплетну врсту из података само када су нам за анализу неопходне све њене ћелије, а иначе не. Размотрићемо процедуру на једном типичном примеру.

Пример 1.3 (коефицијент корелације)**.** Нека су подаци дати матрицом

$$[X, Y] = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

и нека у њој постоје ћелије које недостају. Нека желимо да на основу података оценимо коефицијент корелације $\rho(X, Y)$. Да су сви подаци доступни, то бисмо урадили користећи формулу

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}}.$$

Пар-по-пар приступ ће тада $\overline{x}_n$ рачунати на основу свих доступних података у првој колони из података, а, аналогно, $\overline{y}_n$ ће рачунати на основу свих доступних података из друге колоне. Аналогно ће рачунати и обе суме у имениоцу. Међутим, биће у стању да срачуна i-ти производ из суме у бројиоцу ако и само ако су обе ћелије у i-тој опсервацији доступне. Дакле, за коваријацију из бројиоца се, суштински, ради уклањање свих некомплетних опсервација.

Видимо да су бројилац и именилац суштински рачунати на различитим подацима. То може да доведе до тога да оцена коефицијента корелације испадне из интервала $[-1, 1]$, што је, наравно, бесмислено.

Овај пример нам говори да ни анализа пар-по-пар, иако наизглед унапређење у односу на уклањање некомплетних опсервација, не представља баш паметан начин за руковање недостајућим подацима.

Следећи корак у правом смеру било би уочити да променљиве у подацима углавном нису потпуно независне. За линеарну регресију некорелисаност, или барем слаба корелисаност променљивих које су предиктори, јесте пожељна особина, али је и тада нека функционална зависност предиктора честа појава (рецимо модел с интеракцијама[5]). Све ово говори нам о томе да је веома вероватно да доступни подаци дају неку информацију о онима који недостају, те да би се та информација могла искористити да се недостајући податак **замени неком смисленом вредношћу**. Овај процес познат је под називом *импутација*[6] *недостајућих података* (енг. *Missing Data Imputation*).

О импутацији ће бити више речи у даљем тексту, а сада ћемо се детаљније позабавити самим механизмима недостајања података.

1.4 Класификација недостајања

Као што смо видели из примера 1.2, подаци не морају да недостају на случајан начин. Пример потпуно случајног недостајања били би резултати мерења инструментом који се повремено квари. Дакле, видимо да међу самим начинима на које подаци недостају постоје значајне разлике, па бисмо волели да их на известан начин класификујемо.

Дефиниција 1.1. Нека имамо податке смештене у $n \times K$ матрицу $\mathbf{Y} = [y_{ij}]$. *Матрица одговора*[7] $\mathbf{R} = [r_{ij}]$, истих димензија као и $\mathbf{Y}$, дефинише се као:

$$(\forall i, j)\ r_{ij} = 0 \text{ ако } y_{ij} \text{ недостаје, а 1 иначе.}$$

[4]Иако може бити више од два атрибута, биће јасно из контекста.

[5]Видети [2].

[6]Српски термин био би *попуњавање*, али се у домаћој науци није одомаћио.

[7]Име је потекло из анкета: недостајући податак се појављује тако што неко одбије да одговори на питање. Среће се назив *матрица одзива*.

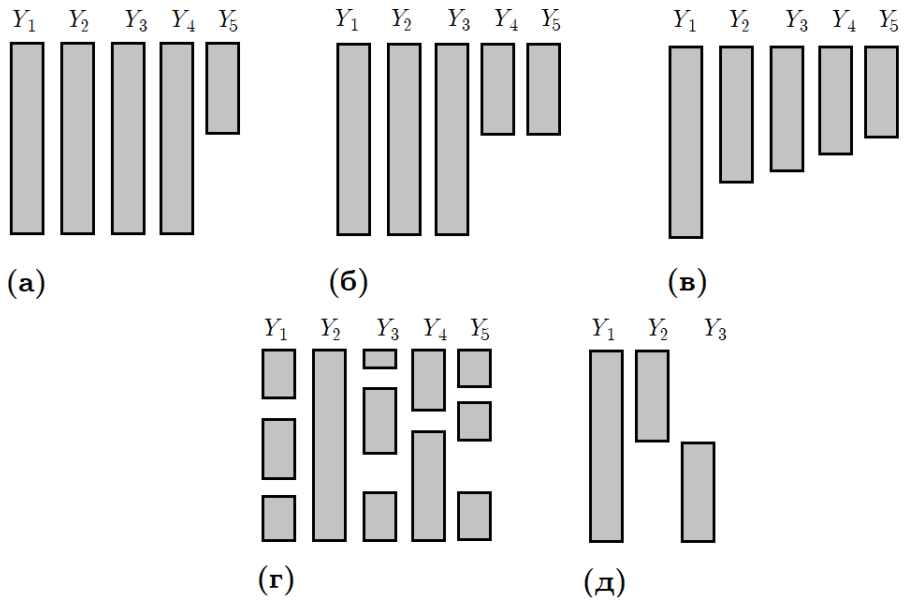
Напомена. Матрица $\mathbf{R}$, дакле, представља извесну „маску" преко матрице $\mathbf{Y}$ и свака њена ћелија је индикатор који постаје 0 ако одговарајућа ћелија у матрици $\mathbf{Y}$ недостаје. Често коришћена алтернатива матрици $\mathbf{R}$ јесте такозвана *индикаторска матрица недостајања* $\mathbf{M}$, са пољем m_{ij} које је 1 ако податак недостаје, а 0 ако је доступан. Она у неку руку представља логичку супротност матрице $\mathbf{R}$ и оне једна другу једнозначно одређују. Ми ћемо се надаље држати матрице $\mathbf{R}$.

Напомена. Понекад се, због разликовања порекла недостајања, у матрици $\mathbf{M}$ (када се она користи) нула користи за доступно поље, а недостајућа поља се кодирају различитим ненула бројевима. Рецимо, у медицинском истраживању, 1 може да значи да је пацијент одустао од истраживања, 2 да је преминуо у току истраживања итд. Ми ћемо се фокусирати на бинарни случај, а о осталима више се може наћи у [5].

1.4.1 Обрасци недостајања

Згодно је разликовати *образац*, или, често, *шаблон*, по којем подаци недостају и разликовати га од *механизма* недостајања. Када кажемо образац недостајања, мислимо на опис недостајања у конкретним подацима које имамо, распоред по колонама, итд, а када кажемо механизам, мислимо на сам вероватносни модел који стоји у позадини недостајућих података. Мислимо, дакле, на матрицу $\mathbf{R}$, њену расподелу, однос са доступним и недоступним подацима. Сумирано, можемо рећи да матрица $\mathbf{R}$ посматрана као случајна величина дефинише један механизам недостајања, а да је свака њена реализација један његов образац.

Пример 1.4 (обрасци недостајања)**.** У овом примеру поменућемо неке честе обрасце недостајања, а више и детаљније може се прочитати у [5], где су дати и примери са реалним подацима. Да бисмо их боље разумели, служићемо се сликом 1.1.



Слика 1.1: Чести обрасци недостајања

(а) **Једнодимензионо недостајање**. На слици 1.1а дат је пример једнодимензионих недостајућих података[8]. У овом случају, недостајање је присутно у само једној колони/променљивој међу подацима. Ово је први образац недостајања који је био научно обрађиван. Честа је појава у ситуацијама у којим је оштећена променљива једина која је резултат неког мерења.

[8]Енг. *Univariate Missing Data.*

(б) **Недостајање истих променљивих у више опсервација**. Ово представља уопштење претходне ситуације, где недостајања могу бити присутна у више променљивих, али на специфичан начин, који је илустрован на слици 1.1б. Уколико у i-тој опсервацији недостаје y_{4i}, недостајаће и y_{5i}. Другим речима, неке променљиве недостају заједно. Ово је честа појава у анкетама, где оштећене променљиве представљају одговоре на „шкакљива" питања.

(в) **Монотоно недостајање.** Оно је присутно у ситуацијама где свака следећа опсервација[9] има све више недостајућих ћелија од претходне. Ово је честа ситуација у медицинским истраживањима, где пацијенти временом одустају од неких (или свих) тестова, јер им постају болни, непријатни итд.

(г) На слици 1.1д дато је уопштење сва три претходна, где се не може уочити никаква правилност. Подаци који прате овакав образац недостајања често приближно прате неки други, правилнији.

(д) **Две променљиве које се не могу заједно опажати.**[10] Када су доступне велике количине података, честа је појава да постоје променљиве, назовимо их Y_2 и Y_3 које за велики број опсервација испољавају тај проблем да ако је доступна једна, друга није. Екстремни случај оваквог обрасца приказан је на слици 1.1д. Основни проблем оваквог обрасца недостајања јесте да се не могу оцењивати параметри међусобне зависности променљивих Y_2 и Y_3.

1.4.2 Механизми недостајања

Нека нам $\mathbf{Y}$ означава податке, овај пут посматране као случајну величину. Уколико међу подацима постоји само један атрибут, тада је $\mathbf{Y}$ случајан колона-вектор. Уколико међу подацима имамо K атрибута, ми ћемо $\mathbf{Y}$ замишљати као $n \times K$ матрицу, али ћемо га често теоријски посматрати као случајан вектор дужине nK, који настаје тако што се матрица „разлиста" идући по врстама. Ово нам неће претерано сметати, јер знамо да су простори $\mathbb{M}_{r\times s}(\mathbb{R})$ и $\mathbb{R}^{rs}$ алгебарски и тополошки изоморфни,[11] а бенефити овакве дуалности веома брзо ће се уочити. Податке $\mathbf{Y}$ увек ће пратити случајни вектор/матрица одговора $\mathbf{R}$. Нотација коју ћемо даље користити делимично је преузета из [5], а делимично из [10].

Напомена. Привремено ћемо се ослободити претпоставке о томе да су подаци реализација n IID случајних вектора. У пракси је углавном тако, али ћемо ипак увести минималан број претпоставки који нам је за излагање теорије неопходан.

Функција[12] $o(\mathbf{Y}, \mathbf{R})$ представљаће нам „подвектор" од $\mathbf{Y}$ који чине они његови елементи који су на позицијама на којима у $\mathbf{R}$ стоји 1. Констатујмо да је $o(\mathbf{Y}, \mathbf{R})$ случајна величина, као функција случајних величина.[13] Када будемо посматрали реализације $\mathbf{Y}$ и $\mathbf{R}$ користићемо ознаке $\widetilde{\mathbf{y}}$ и $\widetilde{\mathbf{r}}$, по потреби са одговарајућим индексима. Означимо са $\overline{o}(\mathbf{Y}, \mathbf{R})$ „подвектор" од $\mathbf{Y}$ који се састоји од оних поља у $\mathbf{Y}$ која су недостајућа. Дакле, $\overline{o}(\mathbf{Y}, \mathbf{R})$ садржи недостајуће податке.

Напомена. Не треба мешати *доступне* и *реализоване* податке. Доступни подаци $o(\mathbf{Y}, \mathbf{R})$ представљају случајну величину (вектор) и имају реализацију $o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})$.

Сада се види и предност дуалности матрице и вектора: када се из вектора избаце недостајућа поља и када се он „скрати" - он остаје вектор - елемент од $\mathbb{R}^d$; није јасно како матрици избацити (неправилно распоређена) недостајућа поља, а задржати структуру матрице. На слици 1.2 илустрована је употреба уведених ознака и поменуте дуалности на конкретном примеру. Сивом бојом означена су поља која недостају, а белом доступна.

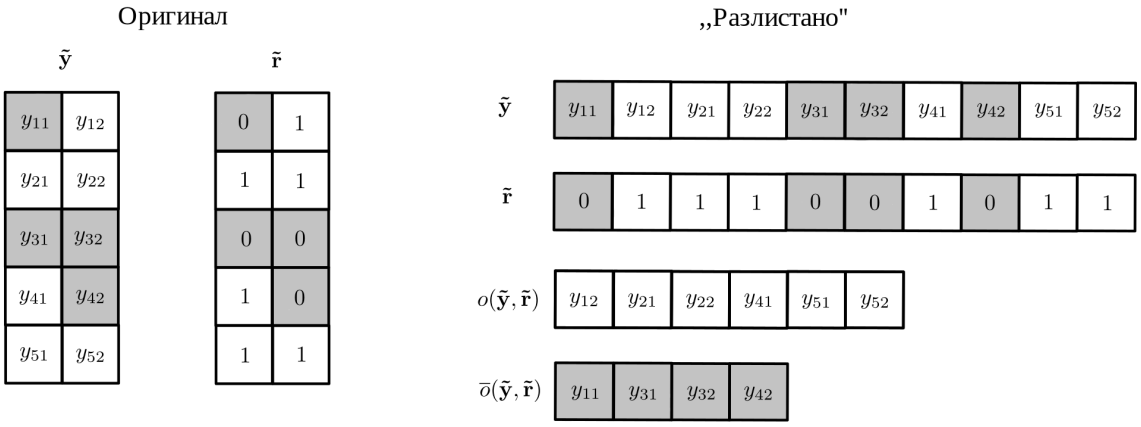
[9] Па самим тим и атрибут, јер се они тако могу поређати.

[10] Енг. *The File-Matching Problem.*

[11] Више о овоме може се прочитати у [13].

[12] Ознака o је од енглеског *observed*: опажен, доступан. Ова ознака је из [10] и згодна је за коришћење у тренутку увођења дефиниције. Литл и Рубин користе $\mathbf{Y}_{(0)}$, а Бурен $\mathbf{Y}_{\text{obs}}$. Ми ћемо, по потреби, користити све ове ознаке, у зависности од контекста.

[13] Нећемо се превише оптерећивати доказом да је Борелова; сматрамо да је то довољно очигледно.



Слика 1.2: Илустрација уведених ознака

Дефиниција 1.2. У досадашњим ознакама, под *моделом (механизмом) недостајања* подразумевамо условну расподелу $\mathbf{R}$ при услову $\mathbf{Y}$, коју ћемо дати њеним законом расподеле

$$g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}) = \mathbf{P}\{\mathbf{R} = \mathbf{r} \mid \mathbf{Y} = \mathbf{y}\}, \tag{1.1}$$

где $\boldsymbol{\phi}$ представља непознате параметре модела.

Сада смо спремни да изложимо Рубинову ([31]) дефиницију случајног недостајања података. Подразумевамо досадашње ознаке.

Дефиниција 1.3. За податке кажемо да *недостају на случајан начин у реализованом смислу*[14] уколико је испуњено:

$$\big(\forall \boldsymbol{\phi}\big)\big(\forall \mathbf{y} :\ o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\big)\ \ g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) = g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}), \tag{1.2}$$

где $\mathbf{y}$ представља реализовану вредност од $\mathbf{Y}$.

Претходна дефиниција нам каже да модел недостајања $g_{\boldsymbol{\phi}}(\cdot \mid \cdot)$ за све вредности параметара $\boldsymbol{\phi}$ претпоставља да условни закон расподеле да образац недостајања $\mathbf{R}$ узме *реализовану* вредност $\widetilde{\mathbf{r}}$ не мења своју вредност, докле год се подаци који су у услову поклапају тамо где су доступни. Још једноставније: вероватноћа да модел недостајања узме *реализовану* вредност при услову података - не зависи од података који недостају.

Напомена. Уочимо да претходна дефиниција не каже ништа о обрасцима недостајања који су *могући, али се нису реализовали.* Такође, претпоставка коју смо увели односи се на претпостављени модел недостајања $g_{\boldsymbol{\phi}}(\cdot \mid \cdot)$, који не мора нужно одговарати стварности.

Наредна дефиниција представља строжију варијанту претходне.

Дефиниција 1.4. Подаци $\mathbf{Y}$ недостају *увек на случајан начин*[15] уколико важи да је

$$\big(\forall \boldsymbol{\phi}\big)\big(\forall\, \mathbf{r}, \mathbf{y}, \mathbf{y}^* :\ o(\mathbf{y}, \mathbf{r}) = o(\mathbf{y}^*, \mathbf{r})\big)\ \ g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}) = g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}^*),$$

где $\mathbf{y}$ и $\mathbf{y}^*$ представљају пар могућих реализација $\mathbf{Y}$.

Претходна дефиниција каже нам да се вредност модела $g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y})$, при фиксном $\mathbf{r}$ (било реализованом или не), не мења при мењању скупа података у услову, све док се ти скупови података поклапају на местима које $\mathbf{r}$ препознаје као доступна.

[14] Енг. *Realised Missing at Random*, или, скраћено, *Realised* MAR.
[15] Енг. *Everywhere MAR.* Може и: свуда на случајан начин.

Напомена. Овакве називе претходних двају механизама преузели смо из [10]. У литератури се RMAR механизам, често, просто зове MAR, док се за EMAR користи синтагма *Missing Always at Random* и скраћеница MAAR. Код Литла и Рубина среће се други начин именовања ([5]). Ми ћемо се трудити да увек нагласимо на који механизам мислимо.

Напомена. У дефиницији 1.3 фигурисала је реализована вредност $\widetilde{\mathbf{r}}$, а у услову све могуће реализације података $\mathbf{Y}$ које се поклапају са конкретном реализацијом $\widetilde{\mathbf{y}}$ на местима која $\mathbf{r}$ означава јединицом. У дефиницији 1.4 појавили су се сви могући обрасци $\mathbf{r}$ и све могуће реализације података $\mathbf{Y}$.

Непажљивом читаоцу могло би се учинити да смо прескочили „средњу могућност": да посматрамо све могуће $\mathbf{r}$ и у услову све податке који се на доступним позицијама поклапају са $\widetilde{\mathbf{y}}$. Штавише, могли смо посматрати и комбинацију реализованог $\widetilde{\mathbf{r}}$ и произвољних података у услову. Ово прво одмах нема смисла: произвољан образац недостајања не може се комбиновати са конкретним реализованим подацима, јер он може као доступно поље препознати оно које је недоступно. Друго нема смисла јер реч „реализовано" која описује $\widetilde{\mathbf{r}}$ губи смисао кад му се склоне подаци над којима се реализовало. Стога смо горњим двама дефиницијама покрили све случајеве који се могу описати реченицом: *недостајање зависи само од доступних података.*

IID случај

Уочимо да је цела досадашња прича из овог пододељка могла бити испричана и за податке који нису нужно IID. Сада ћемо да видимо шта још можемо добити посматрајући тај конкретан случај. Конкретно, претпостављамо да су подаци дати „разлистаном" случајном матрицом,[16] која сада поприма облик $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T$ и да их прати одговарајући индикатор одзива $\mathbf{R} = (\mathbf{R}_1^T, \ldots, \mathbf{R}_n^T)^T$, као и да су $(\mathbf{Y}_i, \mathbf{R}_i)$ **независни и једнако расподељени**[17] (IID). За свако i дефинишемо $o_1(\mathbf{Y}_i, \mathbf{R}_i)$ као подвектор од $\mathbf{Y}_i$, са оних позиција на којима су у $\mathbf{R}_i$ јединице. То су, дакле, реализовани подаци i-те инстанце. Нека $g_{\boldsymbol{\phi},1}(\mathbf{r}_i \mid \mathbf{y}_i)$ представља хипотетички (претпостављени, приписани) модел недостајања за **једну** инстанцу $(\mathbf{Y}_i, \mathbf{R}_i)$, односно

$$g_{\boldsymbol{\phi},1}(\mathbf{r}_i \mid \mathbf{y}_i) = \mathbf{P}\{\mathbf{R}_i = \mathbf{r}_i \mid \mathbf{Y}_i = \mathbf{y}_i\}.$$

Дефиниција 1.5. Подаци недостају по механизму EMAR уколико

$$\big(\forall i \in \{1, \ldots, n\}\big)\big(\forall \boldsymbol{\phi}\big)\big(\forall \mathbf{y}_i, \mathbf{y}_i^* :\ o_1(\mathbf{y}_i, \mathbf{r}_i) = o_1(\mathbf{y}_i^*, \mathbf{r}_i)\big)\ \ g_{\boldsymbol{\phi},1}(\mathbf{r}_i \mid \mathbf{y}_i) = g_{\boldsymbol{\phi},1}(\mathbf{r}_i \mid \mathbf{y}_i^*). \tag{1.3}$$

Лема 1.1. *Ако су $(\mathbf{Y}_i, \mathbf{R}_i)$, $i \in \{1, \ldots, n\}$, независни и једнако расподељени, онда су дефиниције 1.4 и 1.5 еквивалентне.*

Доказ. Тривијално из чињенице да је једна од карактеризација независности та да је производ закона расподеле једнак заједничком закону.

■

Напомена. Лема 1.1 није еквиваленција, већ импликација.

Потпуно случајно недостајање

До сада смо се срели са две врсте података који недостају на случајан начин, и видели смо да се такав механизам недостајања карактерише тиме што расподела недостајања[18] зависи само од доступних података, а не и од оних који недостају. Још слободнији механизам недостајања био би онај у којем расподела недостајања уопште не зависи од података - ни од доступних, ни од недоступних. То је оно што ћемо сада дефинисати.

[16] Матрица која је претворена у вектор, узимајући елементе по врстама.

[17] Ово је веома честа претпоставка, и пракса је показала да је није нереално очекивати.

[18] Ми често помињемо расподелу недостајања, а заправо моделујемо расподелу пристуности. Надамо се да је довољно очигледно да једна другу једнозначно одређује.

Дефиниција 1.6. Кажемо да подаци $\mathbf{Y}$ недостају *на* ***потпуно*** *случајан начин у реализованом смислу*[19] уколико је

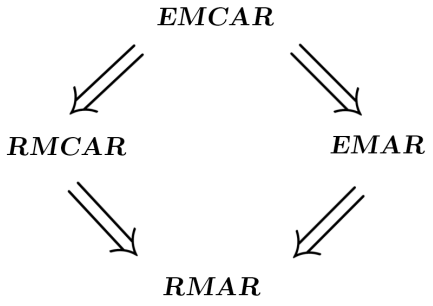
$$(\forall \boldsymbol{\phi})(\forall \mathbf{y}, \mathbf{y}^*)\ \ g_{\boldsymbol{\phi}}(\tilde{\mathbf{r}} \mid \mathbf{y}) = g_{\boldsymbol{\phi}}(\tilde{\mathbf{r}} \mid \mathbf{y}^*). \tag{1.4}$$

Дефиниција 1.7. Кажемо да подаци $\mathbf{Y}$ недостају *увек на* ***потпуно*** *случајан начин*[20] уколико је

$$(\forall \boldsymbol{\phi})(\forall \mathbf{r}, \mathbf{y}, \mathbf{y}^*)\ \ g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}) = g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}^*). \tag{1.5}$$

Слично се могу дефинисати RMCAR и EMCAR механизми за када су $(\mathbf{Y}_i, \mathbf{R}_i)$ IID, и доказати еквиваленција с горњим дефиницијама. Такође, горње дефиниције су у потпуној аналогији са дефиницијама RMAR и EMAR недостајања, те их нећемо дотално појашњавати.

Напомена. Јасно је да EMCAR повлачи RMCAR, EMAR и RMAR, као и да EMAR повлачи RMAR. RMCAR заиста повлачи RMAR, али не и EMAR, што није тешко уочити. Ове односе међу дефинисаним механизмима недостајања лако ћемо запамтити на основу слике 1.3.



Слика 1.3: Односи међу различитим механизмима недостајања

Остало је још да дамо дефиницију недостајања које није случајно.

Дефиниција 1.8. Уколико расподела недостајања (или, јасно, присутности) података $\mathbf{Y}$ зависи од података који недостају, кажемо да је недостајање *неслучајно* и за њега користимо скраћеницу MNAR.[21]

Пример 1.5 (ПСУ са недостајањем)**.** Најједноставнији пример узорка са недостајућим подацима јесте једнодимензиони прост случајан узорак. Нека $\mathbf{Y}$ овде означава цео узорак у смислу случајне величине, као колона-вектор, а слично и $\mathbf{R}$. Нека су $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$ и $\tilde{\mathbf{r}} = (\tilde{r}_1, \ldots, \tilde{r}_n)^T$ опажене реализације. Нека је $f_{\boldsymbol{\theta}}(\cdot)$ густина за Y (у смислу једне компоненте од $\mathbf{Y}$) која зависи од својих (непознатих) параметара $\boldsymbol{\theta}$. Нека су и овде $\boldsymbol{\phi}$ параметри модела недостајања. Претпоставимо још и да су (Y_i, R_i), $i \in \{1, \ldots, n\}$, независни и једнако расподељени [22]. Имамо да је

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\phi}}\{\mathbf{Y} = \tilde{\mathbf{y}}, \mathbf{R} = \tilde{\mathbf{r}}\} &= \mathbf{P}_{\boldsymbol{\theta},\boldsymbol{\phi}}\{\mathbf{R} = (\tilde{r}_1, \ldots, \tilde{r}_n)^T, \mathbf{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\} \\
&= \mathbf{P}_{\boldsymbol{\phi}}\{\mathbf{R} = (\tilde{r}_1, \ldots, \tilde{r}_n)^T \mid \mathbf{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\}\mathbf{P}_{\boldsymbol{\theta}}\{\mathbf{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\} \\
&= \mathbf{P}_{\boldsymbol{\phi}}\{\mathbf{R} = (\tilde{r}_1, \ldots, \tilde{r}_n)^T \mid \mathbf{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\} \prod_{j=1}^{n} f_{\boldsymbol{\theta}}(\tilde{y}_j)
\end{aligned}$$

[19] Користићемо скраћеницу RMCAR, од енглеског *Realised Missing Completely at Random.*

[20] Користићемо скраћеницу EMCAR, од енглеског *Everywhere Missing Completely at Random.* Среће се и скраћеница MACAR, од енглеског *Missing Always Completely at Random*, што је оригинална Рубинова и Литлова скраћеница.

[21] Од енглеског *Missing not at Random.* Ово је тренутно званична терминологија, док се у ранијим верзијама књиге [5] користило *Not Missing at Random.*

[22] Ово су инстанце дужине 1, па их не пишемо масним словима као, нпр, код дефиниције 1.5.

$$= \prod_{i=1}^{n} \mathbf{P}_{\boldsymbol{\phi}}\{R_i = \tilde{r}_i \mid \mathbf{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\} \prod_{j=1}^{n} f_{\boldsymbol{\theta}}(\tilde{y}_j)$$
$$= \prod_{i=1}^{n} g_{\boldsymbol{\phi}}(\tilde{r}_i \mid \tilde{y}_i) \prod_{j=1}^{n} f_{\boldsymbol{\theta}}(\tilde{y}_j),$$

где последња једнакост важи због претпоставке да су (Y_i, R_i) IID.

Уочимо да $g_{\boldsymbol{\phi}}(\tilde{r}_i \mid \tilde{y}_i)$ представља Бернулијев закон расподеле вероватноћа, са вероватноћом да је $\tilde{y}_i$ доступно једнакој $\mathbf{P}_{\boldsymbol{\phi}}\{\tilde{r}_i = 1 \mid \tilde{y}_i\}$. Такође, како је у овом случају свака опсервација вектор дужине један, то јест скалар, овде немамо разлику између MAR и MCAR недостајања, било реализованог или „увек". Дакле, закључујемо да код једнодимензионог ПСУ недостајање може бити или MCAR, или MNAR.

За још примера препоручује се погледати [5], као и [6].

1.5 Тестови механизма недостајања

Приметимо да смо се у досадашњем раду много пута до сада позвали на RMAR претпоставку, или неку од других претпоставки о типу недостајања. Генерално, за реализоване податке и реализовани образац недостајања, ми бисмо желели да тестирамо од ког је механизма недостајање потекло.

Уколико се сетимо дефиниције RMCAR и RMAR недостајања, сетићемо се да су се оне сводиле на то да ли механизам недостајања зависи и од доступних и од недоступних података, или само од доступних, или ни од једних. У првом случају имали смо неслучајно недостајање, други је био (R)MAR, а трећи (R)MCAR. Ми, дакле, треба да тестирамо некакву зависност.

Посматрајући тако ствари, лако закључујемо да је (R)MAR „нетестабилна" претпоставка - није ју могуће тестирати. Зашто? Па просто - подаци који су нам неопходни да то урадимо нису доступни! Слично је за MNAR. Ипак, постоје тестови који тестирају да ли подаци недостају на потпуно случајан начин.

1.5.1 Литлов MCAR тест

Најпознатији тест за тестирање RMCAR[23] јесте такозвани *Литлов MCAR тест*, а предложио га је Родерик J.A. Литл у [27].

Литлов тест почива на јакој претпоставци да је свака врста $\mathbf{y}_i$, $i = 1, \ldots, n$ узоркована из K-димензионе нормалне $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ расподеле. Свакој врсти $\mathbf{y}_i$ одговара образац недостајања (тј. доступности) $\mathbf{r}_i$. Неки од тих образаца могу да се понављају, па претпоставимо да их је укупно J. Претпоставимо да су индекси доступних података j-те инстанце означени са $\mathbf{o}_j$, а недоступних са $\mathbf{m}_j$. Даље, нека је $p_j = |\mathbf{o}_j|$, нека је $\boldsymbol{\mu}_{\mathbf{o}_j}$ средња вредност доступних ћелија j-тог обрасца $(p_j \times 1)$, а нека је $\boldsymbol{\Sigma}_{\mathbf{o}_j}$ одговарајућа $p_j \times p_j$ коваријациона матрица. Нека је $\overline{\mathbf{y}}_j$ $(p_j \times 1)$ одговарајући узорачки просек свих инстанци са j-тим обрасцем, на доступним местима. Коначно, нека је $\mathbf{I}_j \subseteq \{1, 2, \ldots, n\}$ скуп индекса инстанци којима одговара образац $\mathbf{r}_j$ и нека је $n_j = |\mathbf{I}_j|$.

Литлова χ^2 тест статистика дата је изразом

$$d_0^2 = \sum_{j=1}^{J} n_j (\overline{\mathbf{y}}_j - \boldsymbol{\mu}_{\mathbf{o}_j})^T \boldsymbol{\Sigma}_{\mathbf{o}_j}^{-1} (\overline{\mathbf{y}}_j - \boldsymbol{\mu}_{\mathbf{o}_j}). \tag{1.6}$$

Означимо још и са $\mathbf{Y}_{\mathbf{o}_i}$ доступне податке у i-тој инстанци. Уколико је недостајање MCAR, и уколико се сложимо да користимо f за густину/закон расподеле различитих ствари, а да значење разумевамо из контекста, можемо срачунати да је

$$f(\mathbf{y}_{\mathbf{o}_i} \mid \mathbf{r}_i) = \frac{f(\mathbf{y}_{\mathbf{o}_i}, \mathbf{r}_i)}{f(\mathbf{r}_i)} = \frac{f(\mathbf{r}_i \mid \mathbf{y}_{\mathbf{o}_i}) f(\mathbf{y}_{\mathbf{o}_i})}{f(\mathbf{r}_i)} = f(\mathbf{y}_{\mathbf{o}_i}),$$

[23] У време кад је настао није било дистинкције између RMCAR и EMCAR. Када Литл каже MCAR, мисли на RMCAR. Конфузија траје до данас, а ми смо је себи успели разрешити читајући [10].

јер услов може да се склони. Стога, да бисмо одбацили MCAR претпоставку довољно је одбацити хипотезу

$$H_0: \quad [\mathbf{Y}_{\mathbf{o}_i} \mid \mathbf{r}_i] \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{o}_j}, \boldsymbol{\Sigma}_{\mathbf{o}_j}), \qquad i \in \mathbf{I}_j, \;\; 1 \leqslant j \leqslant J \tag{1.7}$$

где је $\boldsymbol{\mu}_{\mathbf{o}_j}$ подвектор од $\boldsymbol{\mu}$, у корист алтернативе

$$H_1: \quad [\mathbf{Y}_{\mathbf{o}_i} \mid \mathbf{r}_i] \sim \mathcal{N}(\boldsymbol{\nu}_{\mathbf{o}_j}, \boldsymbol{\Sigma}_{\mathbf{o}_j}), \qquad i \in \mathbf{I}_j, \;\; 1 \leqslant j \leqslant J \tag{1.8}$$

где $\boldsymbol{\nu}_{\mathbf{o}_j}$ варира у зависности од j.

Литл је у [27] показао да је d_0^2 статистика теста количника веродостојности за тестирање MCAR против H_1. При претпоставци нормалности, расподела јој је χ^2, где је број степени слободе једнак

$$\sum_{j=1}^{J} p_j - K.$$

Ако врсте података нису узорковане из нормалне расподеле, али јесу IID из расподеле са очекивањем $\boldsymbol{\mu}$ и коваријационом матрицом $\boldsymbol{\Sigma}$, онда расподела од d_0^2 важи асимптотски. Уколико се ови параметри не знају, Литл предлаже да се замене својим оценама максималне веродостојности и показује да асимптотска расподела тест статистике остаје иста. Извођења се могу наћи у [27]. Наравно, мисли се на расподелу под нултом хипотезом.

Напомена. Још увек не знамо како да изведемо оцену максималне веродостојности параметра на основу некомплетних података, али ћемо и то сазнати до краја рада.

1.6 Груба класификација метода

У овом одељку понудићемо грубу класификацију метода за рад са недостајућим подацима, у зависности од тога на чему они почивају. Огроман је број таквих метода, и амбиција овог рада никако није да их све изложи. Штавише, аутору није позната ниједна монографија или било какав други текст који их све обухвата. Најприближније томе што постоји јесу [5] и [6]. Напоменимо да класе метода које ћемо у наставку набројати не морају бити дисјунктне, то јест за неке методе није јасно у коју их класу сврстати.

1. **Уклањање некомплетних опсервација.** О овом методу смо, заједно с методом пар-по-пар, говорили на самом почетку, где смо истакли да он, осим у MCAR случају, даје веома пристрасне оцене параметара и води ка погрешном статистичком закључивању. Он се генерално никада не препоручује за употребу, осим ако је удео недостајућих података занемарљиво мали, а механизам засигурно MCAR.

2. **Тежински методи.** Ова група метода почива на томе да се свакој опсервацији придружи извесна *тежина*, да се оцене параметара модификују да укључе те тежине, које треба да компензују присуство недостајућих података. Демонстрираћемо на примеру, узетом из [5].

   Пример 1.6 (HK оцена)**.** У теорији узорака честе су ситуације где се из популације обима N вади узорак обима n, а зна се да је вероватноћа укључења i-те јединке у узорак једнака неком познатом π_i. Тада се често користи Хајекова (1971) оцена популацијске средине:

   $$\overline{y}_{HK} = \frac{\sum_{i=1}^{n} \frac{1}{\pi_i} y_i}{\sum_{i=1}^{n} \frac{1}{\pi_i}},$$

   где сума иде по n јединки из узорка, а не по првих n из популације, што је стандардна „злоупотреба нотације" у теорији узорака.

   Нека је $\hat{p}_i$ оцена вероватноће да i-та јединка **не** недостаје. Она се обично добија тако што се за њу узме удео доступних јединки из неке поткласе узорка у којој се налази i-та

јединка (уколико је доступна стратификација, кластеризација итд). Тада модификовану оцену средине добијамо као

$$\frac{\sum_{i=1}^{n} \frac{1}{\pi_i \hat{p}_i} y_i}{\sum_{i=1}^{n} \frac{1}{\pi_i \hat{p}_i}}.$$

Више о оваквим модификацијама може да се прочита у [5], поглавље 3, а више о оценама сличним Хајековој може се прочитати у [12].

3. **Методи засновани на расподели података.** Ови методи почивају на томе да се формира модел у који се уклапају комплетни подаци, а да се онда закључци изводе користећи било оцене максималне веродостојности под тим моделом, било бајесовско закључивање или нешто треће. О оваквим методима говорићемо у наредној глави, која ће се позабавити тиме када је довољно знати расподелу података, а када (под којим претпоставкама) се мора знати и расподела самог недостајања.

4. **Импутација.** Импутације смо се кратко дотакли на почетку, где смо поменули да она представља процес попуњавања празне (недостајуће) ћелије у подацима неком смисленом вредношћу. Јасно је да уколико се импутиране вредности буду превише разликовале од правих (недоступних), утолико ће и статистичко закључивање мање одговарати стварности. Због тога је веома битно контролисати валидност импутационог метода који се користи, а за то су, донекле, развијене технике.

   Импутационих метода има много и њима ћемо у наставку посветити значајан део овог текста, јер импутација представља убедљиво најкоришћенији приступ за руковање недостајућим подацима.

Глава 2

Приступи засновани на расподели

Циљ ове главе биће да дођемо до закључка у којим се то ситуацијама, тј. при којим претпоставкама, механизам недостајања података може занемарити а да закључци и даље остану валидни, ако претпоставимо да знамо расподелу самих података. Формализација реченог вариране од контекста до контекста, те ће нас она сачекати нешто касније.

2.1 Три начина статистичког закључивања

У овом одељку угрубо ћемо изложити на које начине се све могу изводити статистички закључци на основу података. Јасно, статистички закључци се углавном изводе о вредностима параметара (макар и функционалних), интервалима поверења за те параметре и тестовима везаним за њих. За овакву класификацију, као и дискусију њених односа са недостајућим подацима, захваљујемо се ауторима рада [10]. Три типа статистичког закључивања јесу *закључивање директно на основу веродостојности*, *бајесовско закључивање* и *фреквенционистичко закључивање*. У пресеку првог и трећег налази се *фреквенционистичко закључивање на основу веродостојности*. О свакоме од њих рећи ћемо по нешто, изложивши их **на примеру закључивања без присуства недостајућих података**, док ћемо убрзо дати и то уопштење.

2.1.1 Закључивање директно на основу веродостојности

Код овог вида закључивања претпоставља се да сами подаци $\mathbf{Y}$ имају придружену расподелу, која зависи од коначног скупа (вектора) непознатих параметара $\boldsymbol{\theta}$. Неки од ових параметара јесу од интереса да се о њима изводе закључци, док су други сметајући и ту су због саме природе расподеле. Уочимо ли да заправо расподела података зависи од саме вредности потенцијалне реализације $\mathbf{y}$ и од параметара $\boldsymbol{\theta}$, то је можемо посматрати (или неки њен умножак, у зависности од ситуације) као функцију параметара за фиксне податке, и звати функцијом веродостојности.

Код директног закључивања на основу веродостојности за тачкасту оцену параметра узима се она вредност[1] $\boldsymbol{\theta}$ која максимизује функцију веродостојности. За тестирање се углавном користи количник вредности функције веродостојности у различитим параметрима и сличне модификације.[2] Код оваквог закључивања не говоримо о интервалима поверења, већ о *интервалима веродостојности*, који представљају све вредности параметра за које је функција веродостојности прекорачила унапред задати праг.

Уколико неке од параметара желимо да елиминишемо, позната су два начина. Први је тзв. *профилисани метод веродостојности*. Претпоставимо, без умањења општости, да се па-

[1] Или једна од, ако није јединствена.

[2] Добар пример таквог теста јесте и тест количника веродостојности (LRT - *Likelihood Ratio test*), који је за специјалне типове хипотеза чак и униформно најмоћнији. За детаље можемо видети [18].

раметри могу записати као $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$ и да желимо да елиминишемо $\boldsymbol{\theta}_2$. Тада се тзв. профилисана веродостојност за θ_1 дефинише као $\mathcal{L}(\theta_1, \hat{\boldsymbol{\theta}}_2)$, где је $\mathcal{L}$ функција веродостојности, а $\hat{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}_2} \mathcal{L}(\theta_1, \boldsymbol{\theta}_2)$. Максимум, наравно, иде по свим могућим $\boldsymbol{\theta}_2$ при фиксном θ_1. Алтернатива овом методу јесте тзв. *условна вероdостојност*, где се параметри елиминишу тако што се функција веродостојности не формира од расподеле за $\mathbf{Y}$, већ њене условне расподеле, при неком услову $h(\mathbf{Y})$, тако да та условна расподела има мање параметара. Не постоје теоријске основе на основу којих би се закључило који од ова два начина елиминације параметара је бољи.

2.1.2 Бајесовско закључивање

Код бајесовског закључивања, несигурност о правој вредности параметра $\boldsymbol{\theta}$ коју имамо урачунавамо тако што параметрима придружујемо расподелу вероватноћа, априорну расподелу $\pi(\boldsymbol{\theta})$ (коју ћемо, БУО, замишљати као густину[3]) коју затим „ажурирамо" подацима добијајући тако тзв. апостериорну расподелу параметара $\boldsymbol{\theta}$ при реализованим подацима $\tilde{\mathbf{y}}$:

$$\pi(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}) = \frac{f_{\boldsymbol{\theta}}(\tilde{\mathbf{y}})\pi(\boldsymbol{\theta})}{\int\limits_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\tilde{\mathbf{y}})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}, \tag{2.1}$$

где је $f_{\boldsymbol{\theta}}$ густина/закон расподеле података $\mathbf{Y}$, а $\boldsymbol{\Theta}$ параметарски простор. Очекивање апостериорне расподеле углавном се узима за оцену параметра. Наравно, ово се може уопштити. Дефинише се тзв. *функција губитака* $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ која треба да „мери" разлику између стварне и оцењене вредности параметра (веће L - већи губитак), те на основу ње да се дефинише *апостериорни средњи ризик оцене* као очекивани губитак, где очекивање иде по апостериорној расподели параметра:

$$R(\hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{\Theta}} L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \tilde{\mathbf{y}})\mathrm{d}\boldsymbol{\theta}.$$

Сада се за оптималну оцену узима она која минимизује апостериорни средњи ризик. За једнодимензиони параметар се може показати да ако се узме квадратна[4] функција губитака, добије се баш апостериорно очекивање као оптимална оцена. Ако се узме апсолутна функција губитака, за оцену се добије медијана апостериорне расподеле, а ако је $L(\hat{\theta}, \theta) = I\{\hat{\theta} \neq \theta\}$ онда је оцена мода расподеле. О овоме се више може прочитати у [16].

Како је код бајесовског закључивања параметар (претпоставимо једнодимензион) случајна величина, то можемо рачунати вероватноћу да он упадне у неки фиксан, неслучајан реални интервал. Такве интервале, да бисмо их разликовали од интервала поверења, називамо *интервалима прекривања*, и углавном их добијамо као интервале између два квантила апостериорне расподеле.

2.1.3 Фреквенционистичко закључивање

Приметимо да је код оба претходно наведена вида закључивања заједничко то да зависе само од реализоване вредности података $\mathbf{Y}$. Нигде се нису узимале у обзир друге могуће, али не реализоване вредности података. Фреквенционистички приступ заузима ту тачку гледишта: њега занима хипотетичко узастопно узорковање података $\mathbf{Y}$ из њихове расподеле као и, одатле исходеће, оцене параметара, затим интервали поверења и слично. Појмови попут пристрасности оцене, њене дисперзије, интервала поверења итд. тек код овог приступа добијају смисао.

Генерално говорећи, код фреквенционистичког приступа бира се функција $t(\mathbf{Y})$ података $\mathbf{Y}$, те се њена реализована вредност $t(\tilde{\mathbf{y}})$ пореди са расподелом за $t(\mathbf{Y})$. Ова расподела, по потреби, може бити и условна при разним условима са реализованим подацима.

[3] Сума ће по потреби постати интеграл и сл.

[4] $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$.

Фреквенционистичко закључивање на основу веродостојности

Након што се формира оцена максималне веродостојности $\hat{\boldsymbol{\theta}}$ чисто на основу тог приступа, са реализованих података $\widetilde{\mathbf{y}}$ пређе се на њихов случајни аналогон, и оцена се посматра као фреквенционистичка оцена $\hat{\boldsymbol{\theta}}(\mathbf{Y})$, те се онда, у зависности од њене димензије, рачуна пристрасност, дисперзија/коваријациона матрица, интервали поверења итд. и даље се наставља као са фреквенционистичком оценом. У оваквим приступима честа је примена централне граничне теореме, јер се не мора нужно знати много о расподели оцене, чак иако се зна много о расподели података, јер сама оцена може бити компликована функција података.

Напомена. И за бајесовске оцене важи слично: и њима се могу испитивати фреквенционистичке особине преласком са $\widetilde{\mathbf{y}}$ на $\mathbf{Y}$. Овиме се нећемо детаљније бавити.

2.2 Игнорабилност

У овом одељку ћемо појаснити под којим претпоставкама можемо занемарити механизам недостајања података за сваки од типова статистичког закључивања које смо претходно објаснили. У овом тренутку и даље нећемо формално одредити шта значи да је механизам недостајања „игнорабилан", већ ћемо се задовољити интуитивном представом да то значи да су закључци изведени из параметарског модела за саме податке исти као закључци изведени из заједничког модела за податке и механизам недостајања. Остаје још да дефинишемо шта значи да су „закључци исти". О томе ћемо касније рећи више.

Нека имамо заједнички параметарски модел за податке $\mathbf{Y}$ и образац недостајања $\mathbf{R}$. Заједничка густина (или закон) за $\mathbf{Y}$ и $\mathbf{R}$ је тада дата са

$$f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y}),$$

где је $f_{\boldsymbol{\theta}}(\cdot)$ густина или закон расподеле података $\mathbf{Y}$. Генерално није битно да ли говоримо о дискретној или непрекидној расподели података, с тим што код непрекидне расподеле имплицитно претпостављамо да сваки интеграл који се појављује - конвергира. Такође, за дискретну расподелу података сваки интеграл треба заменити сумом (која, наравно, мора бити коначна). Све ово ће, надамо се, бити јасно из конкретног контекста.

Нека је $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ параметарски простор за $\boldsymbol{\theta}$, а $\boldsymbol{\Omega}_{\boldsymbol{\phi}}$ параметарски простор за $\boldsymbol{\phi}$. Нека $\boldsymbol{\Omega}_{\boldsymbol{\theta},\boldsymbol{\phi}}$ представља параметарски простор за $(\boldsymbol{\theta}, \boldsymbol{\phi})$ и нека су, као и до сад, $\widetilde{\mathbf{y}}$ и $\widetilde{\mathbf{r}}$ дате реализације од $\mathbf{Y}$ и $\mathbf{R}$. Следи неколико дефиниција које се тичу веродостојности.

Дефиниција 2.1. *Заједничка функција веродостојности* за $(\boldsymbol{\theta}, \boldsymbol{\phi})$ јесте функција чији је домен $\boldsymbol{\Omega}_{\boldsymbol{\theta},\boldsymbol{\phi}}$ и која је дата формулом

$$\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi}) = \int f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y})I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{2.2}$$

Напомена. Уочимо да претходни интеграл (или сума у дискретном случају) заправо елиминише недостајуће податке тако што их „проинтеграли".

Дефиниција 2.2. *Функција веродостојности за параметре $\boldsymbol{\theta}$ игноришући механизам недостајања* јесте функција чији је домен $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ и која је дата формулом

$$\mathcal{L}_2(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}(\mathbf{y})I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{2.3}$$

Дефиниција 2.3. За било које фиксно $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$, *функција веродостојности за $\boldsymbol{\theta}$ при фиксном $\boldsymbol{\phi}$* јесте функција са доменом $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ дата са

$$\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta}) = I_{\{(\boldsymbol{\theta},\boldsymbol{\phi})|(\boldsymbol{\theta},\boldsymbol{\phi})\in\boldsymbol{\Omega}_{\boldsymbol{\theta},\boldsymbol{\phi}}\}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \int f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y})I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y}. \tag{2.4}$$

Дефиниција 2.4. *Профилисана функција веродостојности за $\boldsymbol{\theta}$* јесте функција са доменом $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ дата са

$$\mathcal{L}_4(\boldsymbol{\theta}) = \max_{\boldsymbol{\phi}\in\boldsymbol{\Omega}_{\boldsymbol{\phi}}} \mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta}). \tag{2.5}$$

Напомена. Ову функцију дефинишемо, разуме се, онда када максимум постоји. То у пракси најчешће јесте случај.

Сада већ можемо детаљније описати игнорабилност. Њу ћемо дати у контексту дефиниције, иако та дефиниција није до краја математички прецизна.

Дефиниција 2.5. Кажемо да је механизам недостајања *игнорабилан* за извођење закључака о $\boldsymbol{\theta}$, уколико су закључци о $\boldsymbol{\theta}$ исти на основу $\mathcal{L}_1$ и на основу $\mathcal{L}_2$.

Непрецизност горње дефиниције је у томе што нисмо описали шта значи да су „закључци исти". То се и не може дефинисати у општем облику, већ ће варирати у зависности од вида закључивања који се користи[5], што ћемо обрадити детаљније. Горњу дефиницију, стога, треба схватити као „генеричку".

2.2.1 Закључивање директно на основу веродостојности

Пре било какве дискусије формулисаћемо и доказати наредну теорему.[6] Све ознаке су као и до сада.

Теорема 2.1. ***Претпоставимо да подаци недостају према механизму RMAR и да се параметарски простор може записати као $\boldsymbol{\Omega_{\theta,\phi}} = \boldsymbol{\Omega_\theta} \times \boldsymbol{\Omega_\phi}$. Тада:***

1. ***$\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ се може факторисати на производ два чиниоца, тако да један зависи само од $\boldsymbol{\theta}$, а други само од $\boldsymbol{\phi}$;***
2. ***За произвољно $\boldsymbol{\phi} \in \boldsymbol{\Omega_\phi}$ које задовољава да је $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) > 0$, $\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$ је пропорционално са $\mathcal{L}_2(\boldsymbol{\theta})$;***
3. ***Ако постоји $\boldsymbol{\phi} \in \boldsymbol{\Omega_\phi}$ такво да је $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) > 0$, онда је $\mathcal{L}_4(\boldsymbol{\theta})$ специјалан случај од $\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$, па је стога $\mathcal{L}_4(\boldsymbol{\theta})$ пропорционално са $\mathcal{L}_2(\boldsymbol{\theta})$.***

Доказ. Прво рачунамо да је

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \int f_{\boldsymbol{\theta}}(\mathbf{y}) \underbrace{g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y})}_{\substack{=g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}}|\widetilde{\mathbf{y}}) \\ \text{јер важи RMAR}}} I_{\{\mathbf{y} \mid o(\mathbf{y},\widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\ &= g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) \cdot \int f_{\boldsymbol{\theta}}(\mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y},\widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y},\end{aligned}$$

Како је $\boldsymbol{\Omega_{\theta,\phi}} = \boldsymbol{\Omega_\theta} \times \boldsymbol{\Omega_\phi}$, то значи да кад год $\boldsymbol{\theta} \in \boldsymbol{\Omega_\theta}$ и $\boldsymbol{\phi} \in \boldsymbol{\Omega_\phi}$, тада и $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \boldsymbol{\Omega_{\theta,\phi}}$, па горња једнакост представља жељену факторизацију, чиме смо доказали ставку 1. Такође, одавде можемо извући и да је

$$I_{\{(\boldsymbol{\theta},\boldsymbol{\phi}) \mid (\boldsymbol{\theta},\boldsymbol{\phi}) \in \boldsymbol{\Omega_{\theta,\phi}}\}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1,$$

кад год је $\boldsymbol{\theta} \in \boldsymbol{\Omega_\theta}$ и $\boldsymbol{\phi} \in \boldsymbol{\Omega_\phi}$. Одавде се директно добија да је

$$\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta}) = g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) \cdot \int f_{\boldsymbol{\theta}}(\mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y},\widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y}.$$

Како $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}})$ није функција од $\boldsymbol{\theta}$, то из горње једнакости важи да је $\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$ пропорционално са $\mathcal{L}_2(\boldsymbol{\theta})$ кад год је $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) > 0$, чиме смо доказали и ставку 2.

Такође, због поменутих претпоставки можемо срачунати и да је

$$\mathcal{L}_4(\boldsymbol{\theta}) = \max_{\boldsymbol{\phi} \in \boldsymbol{\Omega_\phi}} \mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$$

[5]На пример, код закључивања директно на основу веродостојности за оцену се узима она вредност која максимизује функцију веродостојности, а код бајесовског закључивања узима се, углавном, очекивање апостериорне расподеле.

[6]Преузели смо је, са доказом, из [10], доказали су је Литл и Рубин, али нисмо успели пронаћи у којем раду и које године. Без доказа се може наћи и у [5]

$$= \max_{\boldsymbol{\phi}\in\boldsymbol{\Omega}_{\boldsymbol{\phi}}} \left(g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) \cdot \int f_{\boldsymbol{\theta}}(\mathbf{y}) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} \right)$$
$$= \left(\int f_{\boldsymbol{\theta}}(\mathbf{y}) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} \right) \cdot \max_{\boldsymbol{\phi}\in\boldsymbol{\Omega}_{\boldsymbol{\phi}}} g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}).$$

Јасно $\max_{\boldsymbol{\phi}\in\boldsymbol{\Omega}_{\boldsymbol{\phi}}} g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}})$ не зависи од $\boldsymbol{\theta}$. Штавише, оно је ненула чим постоји једно $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$ за које је $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) > 0$. Сада је $\mathcal{L}_4(\boldsymbol{\theta}) = \mathcal{L}_{3,\hat{\boldsymbol{\phi}}}(\boldsymbol{\theta})$, где је $\hat{\boldsymbol{\phi}}$ оно за које се постиже горњи максимум. Овиме смо доказали и ставку 3, чиме смо комплетирали доказ теореме. ■

Сада је тренутак да прокоментаришемо смисао горње теореме. Ставка 1. у теореми нам каже да се заједничка веродостојност може максимизовати посебно по $\boldsymbol{\theta}$ и по $\boldsymbol{\phi}$. Суштина тога је да се добије исто $\boldsymbol{\theta}$ кад се максимизује било која од њих, па није битно да ли урачунавамо механизам недостајања у функцију веродостојности. То је, наравно, битно, јер се при овом виду закључивања закључци о параметрима изводе максимизацијом функције веродостојности. Сличне закључке дају нам и ставке 2. и 3. Суштина је дакле, **да се закључци могу изводити само на основу $\mathcal{L}_2$, и неће бити погрешни.**

Напомена. До сада смо се позабавили елиминацијом параметара $\boldsymbol{\phi}$ као сметајућих[7] параметара. Уколико бисмо желели, рецимо, интервал веродостојности за једну компоненту θ_1 од $\boldsymbol{\theta}$, требало би елиминисати остале, које ћемо звати $\boldsymbol{\theta}_2$. Ако $\boldsymbol{\theta}_2$ елиминишемо из $\mathcal{L}_2$ и из $\mathcal{L}_4$ коришћењем профилисане веродостојности, пропорционалност ових функција веродостојности гарантује нам и пропорционалност одговарајућих профилисаних, па ће интервали добијени из обе бити исти.

Геометријска интерпретација

Размотримо модел у којем су параметри једнодимензиони: θ и ϕ. График функције $\mathcal{L}_1(\theta, \phi)$ је тада површ у тродимензионом простору. RMAR претпоставка, која нам је дала пропорционалност $\mathcal{L}_2(\theta)$ и $\mathcal{L}_1(\theta, \phi)$ за фиксно ϕ, што заправо намеће геометријску структуру на ту површ: све криве $\theta \mapsto \mathcal{L}_1(\theta, \phi)$, за разне ϕ представљају скалиране копије једна друге, што у општем случају не би морало да буде тачно.

С друге стране, функција $\mathcal{L}_1$ дефинисана је на $\boldsymbol{\Omega}_{\theta,\phi}$, па у општем случају криве добијене из $\mathcal{L}_1$ за фиксно ϕ неће бити дефинисане на целом $\boldsymbol{\Omega}_{\theta}$, већ неком његовом подскупу. Претпоставка о „раздвојености" параметарских простора отклања и овај проблем.

2.2.2 Бајесовско закључивање

Нека нам $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ означава заједничку априорну расподелу за параметре $(\boldsymbol{\theta}, \boldsymbol{\phi})$ (густина или закон расподеле) и нека нам $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ означава одговарајућу маргиналну априорну расподелу. Кажемо да је механизам недостајања игнорабилан за бајесовско закључивање уколико је маргинална апостериорна расподела за $\boldsymbol{\theta}$ добијена заједничким моделовањем и $\mathbf{Y}$ и $\mathbf{R}$ иста као апостериорна расподела за $\boldsymbol{\theta}$ добијена моделовањем само $\mathbf{Y}$.

Наредна теорема даје нам довољне услове за игнорабилност код бајесовског закључивања.

Теорема 2.2. ***Претпоставимо да подаци недостају у складу са RMAR механизмом и да су параметри $\boldsymbol{\theta}$ и ϕ априорно независни. Тада је апостериорна расподела за $\boldsymbol{\theta}$ добијена коришћењем функције веродостојности $\mathcal{L}_1$ и априорне расподеле $\pi_{\boldsymbol{\theta}}(\theta)$ иста као апостериорна расподела за веродостојност $\mathcal{L}_1(\boldsymbol{\theta}, \phi)$ и априорне расподеле $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \phi)$.***

Доказ. Из формуле (2.1) видимо да је апостериорна расподела пропорционална производу густине/закона и априорне расподеле. Јасно, густина/закон у случају комплетних података не представљају ништа друго до функцију веродостојности, само с акцентом на други аргумент. У случају некомплетних података, апостериорну расподелу добијамо тако што на то место стављамо одговарајућу функцију веродостојности која урачунава недостајање.

Уколико користимо $\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ и априорну расподелу $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi})$, добијамо да је апостериорна расподела пропорционална са $\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\phi})\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi})$. Како су $\boldsymbol{\theta}$ и $\boldsymbol{\phi}$ априорно независни, ово

[7] Енг. *nuisance*.

се своди на $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi})\mathcal{L}_1(\boldsymbol{\theta},\boldsymbol{\phi})$, где је $\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ априорна расподела за $\boldsymbol{\phi}$. Коначно, при претпоставци RMAR недостања, теорема 2.1 нам даје пропорционалност $\mathcal{L}_2(\boldsymbol{\theta})$ и $\mathcal{L}_1(\boldsymbol{\theta},\boldsymbol{\phi})$ при сваком фиксном $\boldsymbol{\phi}$, па на крају добијамо да је, за свако $\boldsymbol{\phi}$,

$$\pi_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{\theta},\boldsymbol{\phi})\mathcal{L}_1(\boldsymbol{\theta},\boldsymbol{\phi}) \propto \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathcal{L}_2(\theta),$$

одакле једнакост апостериорних расподела тривијално следи, додавањем скалирајућих константи које од горњих израза праве густине/законе. ■

2.2.3 Фреквенционистичко закључивање

Подсетимо се, заједничка расподела $\mathbf{Y}$ и $\mathbf{R}$ дата је са $f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\mathbf{r} \mid \mathbf{y})$. За свако $\boldsymbol{\phi}$ за које постоји $\mathbf{y}$ такво да је $f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) > 0$ можемо рачунати условну расподелу од $o(\mathbf{Y},\mathbf{R})$ при услову $\mathbf{R} = \widetilde{\mathbf{r}}$ као

$$\frac{\displaystyle\int f_{\boldsymbol{\theta}}(\mathbf{u})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{u})I_{\{\mathbf{u}|o(\mathbf{u},\widetilde{\mathbf{r}})=o(\mathbf{y},\widetilde{\mathbf{r}})\}}(\mathbf{u})\,\mathrm{d}\mathbf{u}}{\displaystyle\int f_{\boldsymbol{\theta}}(\mathbf{u})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{u})\,\mathrm{d}\mathbf{u}}. \tag{2.6}$$

У општем случају ова расподела ће зависити од $\boldsymbol{\phi}$ и неће бити једнака расподели

$$\int f_{\boldsymbol{\theta}}(\mathbf{u})I_{\{\mathbf{u}|o(\mathbf{u},\widetilde{\mathbf{r}})=o(\mathbf{y},\widetilde{\mathbf{r}})\}}(\mathbf{u})\,\mathrm{d}\mathbf{u}. \tag{2.7}$$

Ипак, јасно је да расподела (2.7) на известан начин одговара функцији веродостојности $\mathcal{L}_2$.

Сада је тренутак да „направимо пресек" и сетимо се шта поредимо. Расподела (2.6) представља густину/закон условне расподеле $[o(\mathbf{Y},\mathbf{R}) \mid \mathbf{R} = \widetilde{\mathbf{r}}]$. Често ће нас занимати расподела неке функције на доступним подацима:

$$T = t\left(o(\mathbf{Y},\mathbf{R}),\mathbf{R}\right).$$

Рубин је расподелу од $T \mid \mathbf{R} = \widetilde{\mathbf{r}}$, која се добија користећи (2.6), звао *исправна условна расподела узорковања*[8] Када бисмо за извођење ове расподеле користили (2.7) као основу[9], добили бисмо *потенцијално неисправну условну расподелу узорковања.*[10] Код фреквенционистичког закључивања рећи ћемо да је механизам недостајања игнорабилан онда када су ове две расподеле једнаке[11]. Довољне услове за игнорабилност гарантује нам наредна теорема.

Теорема 2.3. ***Нека подаци недостају према RMCAR механизму и нека постоји*** $\mathbf{y}$ ***за које је*** $f_{\boldsymbol{\theta}}(\mathbf{y})g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) > 0$***. Тада је потенцијално неисправна расподела узорковања једнака исправној условној расподели узорковања.***

Доказ. Када су подаци RMCAR, $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{u})$ не зависи од $\mathbf{u}$, па може да „изађе" испред интеграла, те се израз (2.6) своди на израз (2.7), што завршава доказ. ■

Напомена (дискусија стваралаца теорије). Уочимо да смо расподелу из теореме 2.3 условљавали помоћу $\mathbf{R} = \widetilde{\mathbf{r}}$. Литл је имао мишљење да је то погрешно, јер $\mathbf{R}$ није другоразредна статистика за $\boldsymbol{\theta}$ осим у случају много јачег EMCAR услова, што може бити проблем, јер ће се изгубити део информације о параметру. Међутим, Рубин је уочио да уобичајена дефиниција другоразредности[12] овде није адекватна, и треба је модификовати, јер наше закључивање зависи само од њене реализоване вредности, а не расподеле у општем случају. Хеитјан је ову дискусију наставио тако што је увео концепт *реализовано-другоразредне статистике*[13] и

[8] Енг. *correct conditional sampling distribution*. Видети [31]. Јако је тешко превести на српски језик; „узорачка расподела" је дефинисано за расподелу самог узорка, па је онда ово најприближније, а да се не изгуби смисао. Истина, звучи вештачки.

[9] У смислу расподеле за $[o(\mathbf{Y},\mathbf{R}) \mid \mathbf{R} = \widetilde{\mathbf{r}}]$.

[10] Хеитјан и Басу, [23].

[11] Што је смислено, јер се све особине оцене (пристрасност, дисперзија итд) код фреквенционистичког приступа изводе из њене расподеле.

[12] Да расподела другоразредне статистике не зависи од параметра за који је она другоразредна.

[13] Енг. *observed ancillary statistic*. Детаљније у [21].

сложио се с Рубином у закључку да је за теорему 2.3 адекватно условљавати са $\mathbf{R} = \tilde{\mathbf{r}}$ кад год је недостајање RMCAR.

Иако смо употребили реч *игнорабилно* када смо описали механизам недостајања, уочимо да он овде није баш у потпуности игнорисан, јер смо све расподеле условљавали његовом реализацијом. Ипак ћемо задржати назив.

Фреквенционистичко закључивање на основу веродостојности

Подсетимо се, ово закључивање је специјалан случај фреквенционистичког закључивања, где се оцена параметра прво добије методом максималне веродостојности (и његовим дериватима); затим се уочи да је таква оцена функција реализоване вредности података, $\mathbf{y}$, те се онда они у том изразу замене подацима као случајном величином, $\mathbf{Y}$, те се добије оцена као случајна величина, која има своју расподелу[14] на основу које изводимо фреквенционистичка својства оцене, попут пристрасности, стандардне грешке итд. Стога закључујемо да и у овом случају важи теорема 2.3.

За фреквенционистичко закључивање на основу веродостојности се, ипак, може добити и више. Уочимо за почетак да када подаци нису RMCAR, $g_{\boldsymbol{\phi}}(\tilde{\mathbf{r}} \mid \mathbf{u})$ не зависи од $\mathbf{u}$ и неће моћи да „изађе" испред интеграла у (2.6), па закључци неће бити исти коришћењем те расподеле и расподеле (2.7). Међутим, при услову EMAR недостајања и „раздвојености параметара" ($\boldsymbol{\Omega}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \boldsymbol{\Omega}_{\boldsymbol{\theta}} \times \boldsymbol{\Omega}_{\boldsymbol{\phi}}$), из теореме 2.1 имамо да су $\mathcal{L}_2(\boldsymbol{\theta})$, $\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$ и $\mathcal{L}_4(\boldsymbol{\theta})$ пропорционални не само за реализовано $\mathbf{y}$, већ за све $\tilde{\mathbf{y}}$ добијене у хипотетичком поновном узорковању (узорковањем података „узоркују" се и нова недостајања, па се мењају реализације и за $\mathbf{Y}$ и за $\mathbf{R}$). Одавде закључујемо да се при раздвојености параметара и EMAR недостајању поклапају оцене методом максималне веродостојности добијене на основу различитих функција веродостојности поклапају за све реализације узорка, па како су исте (као функције реализованих узорака), исте су им и расподеле (када се реализовани узорак замени хипотетичким случајним). Стога се добијају исти фреквенционистички закључци за оцене добијене ММВ коришћењем било $\mathcal{L}_1$, било $\mathcal{L}_2$.

Фреквенционистичко бајесовско закључивање

Као што смо оценама методом максималне веродостојности тражили фреквенционистичка својства тако што смо их посматрали као функције не реализованог, већ хипотетичког случајног узорка, то исто можемо радити и са бајесовским оценама.

И у бајесовском случају, при EMAR претпоставци, а још и априорној независности параметара $\boldsymbol{\theta}$ и $\boldsymbol{\phi}$ добијамо да су одговарајуће функције веродостојности пропорционалне за сваку реализацију $(\mathbf{Y}, \mathbf{R})$, па ће увек давати и исте апостериорне расподеле, те ће и фреквенционистичка својства бајесових оцена добијена на основу њих бити иста.

Сада ћемо све до сада речено у овој глави представити згодним дијаграмом.

На слици 2.1 графички је приказано под којим **довољним** претпоставкама је механизам недостајања игнорабилан за различите видове статистичког закључивања и то:

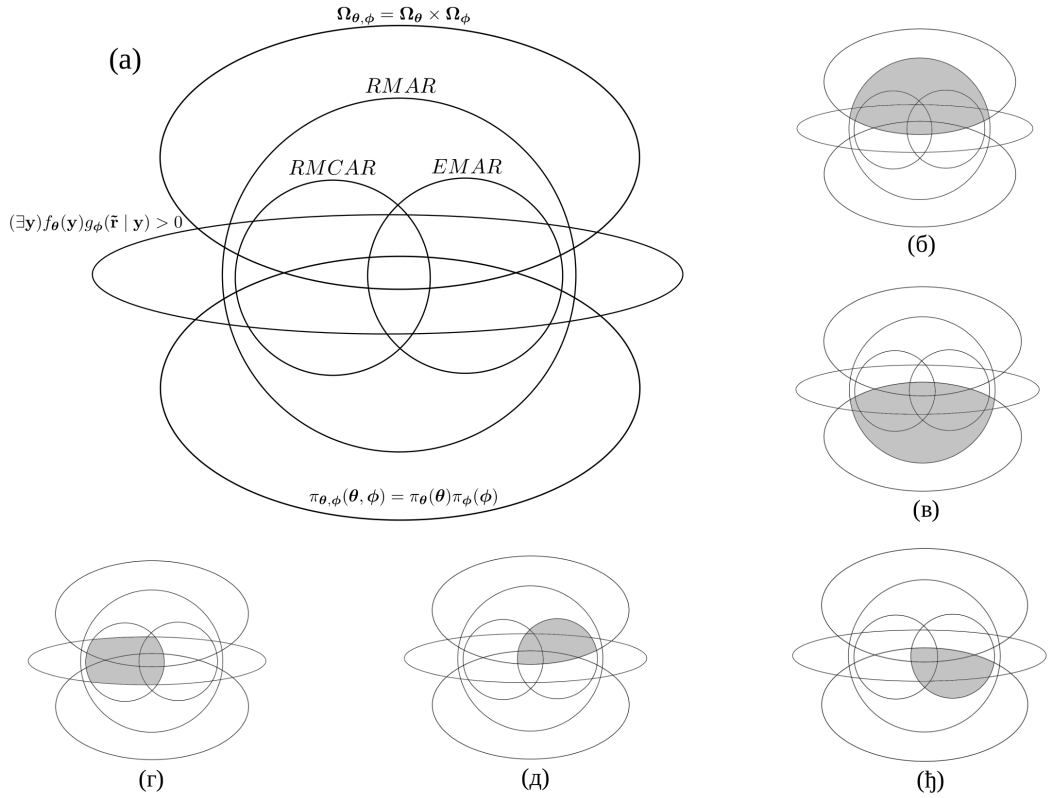
(а) Генералан однос претпоставки; елементи скупова на Веновом дијаграму су заједнички модели за податке и недостајање;

(б) Претпоставке под којима је механизам недостајања ингнорабилан за закључивање директно на основу веродостојности;

(в) Претпоставке под којима је механизам недостајања ингнорабилан за бајесовско закључивање;

(г) Претпоставке под којима је механизам недостајања ингнорабилан за фреквенционистичко закључивање;

[14] Која се на енглеском језику зове *sampling distribution*, а ми смо се одлучили да је зовемо расподела узорковања, како бисмо је разликовали од узорачке расподеле, а то је расподела од $\mathbf{Y}$, што је на енглеском *sample distribution*.



Слика 2.1: Графички приказ игнорабилности при различитим **довољним** претпоставкама

(д) Претпоставке под којима је механизам недостајања ингнорабилан за фреквенционистичко закључивање на основу веродостојности;

(ђ) Претпоставке под којима је механизам недостајања ингнорабилан за фреквенционистичко бајесовско закључивање.

2.3 Условљавање статистиком

У одељцима 2.1.1 и 2.1.3 поменули смо да се често, поред условљавања условом $\mathbf{R} = \tilde{\mathbf{r}}$, врши и условљавање неком функцијом података, $h(\mathbf{Y})$. Један пример за то била је условна веродостојност - начин да се елиминишу сметајући параметри при закључивању методом максималне веродостојности. Сада ћемо ту теорију мало детаљније разрадити.

Нека је $\mathbf{X} = h(\mathbf{Y})$ нека (Борелова, макар и векторска) функција података $\mathbf{Y}$ и нека је $\tilde{\mathbf{x}}$ њена реализована вредност.

Пример 2.1 (опет линеарна регресија)**.** Подаци $\mathbf{Y}$ у линеарној регресији састоје се од циљне променљиве Y и осталих променљивих које зовемо предиктори или атрибути. Нама је циљ извући закључке о Y, па се, заправо, у линеарној регресији све што се моделује - моделује условно под предикторима (што смо видели у примеру 1.1). Ако предикторе означимо са $\mathbf{X}$, јасно је да је $\mathbf{X} = h(\mathbf{Y})$, где функција h представља пројекцију на потпростор атрибута.

Вратимо се на општу причу. Претпоставимо да важи једна од следеће две ствари:

- $\tilde{\mathbf{x}}$ је комплетно (нема недостајућих поља);
- $\int f_{\boldsymbol{\theta}}(y \mid \mathbf{x} = \tilde{\mathbf{x}}) I_{\{\mathbf{y} \mid o(\mathbf{y},\tilde{\mathbf{r}}) = o(\tilde{\mathbf{y}},\tilde{\mathbf{r}})\}}(\mathbf{y})\, \mathrm{d}\mathbf{y}$ не зависи од вредности у $\tilde{\mathbf{x}}$ које недостају, где је $f_{\boldsymbol{\theta}}(y \mid \mathbf{x} = \tilde{\mathbf{x}})$ одговарајућа условна густина/закон.

Када се у дефиниционим изразима за $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ и $\mathcal{L}_4$ израз $f_{\boldsymbol{\theta}}(\mathbf{y})$ замени са $f_{\boldsymbol{\theta}}(y \mid \mathbf{x} = \widetilde{\mathbf{x}})$, ништа суштински се не мења (само је расподела података другачија), па теорема 2.1 и даље важи. Слично, задржаће се и пропорционалност $\mathcal{L}_2$ и $\mathcal{L}_4$ при EMAR претпоставци (где је у обе функције обична густина замењена условном), па пролази и игнорабилност за фреквенционистичко закључивање на основу веродостојности. Уочимо да је при овом, специјалном, случају фреквенционистичког закључивања расподела оцене условна у односу на $\mathbf{X} = \widetilde{\mathbf{x}}$, али не и $\mathbf{R} = \widetilde{\mathbf{r}}$.

Размотримо сада чисто фреквенционистичко закључивање условљено са $\mathbf{X} = \widetilde{\mathbf{x}}$ и $\mathbf{R} = \widetilde{\mathbf{r}}$. Узмимо функцију $T = t\left(o(\mathbf{Y}, \mathbf{R}), \mathbf{R}\right)$ и претпоставимо да је или

- $\widetilde{\mathbf{x}}$ комплетно, или да
- расподела од $[T \mid \mathbf{X} = \widetilde{\mathbf{x}}, \mathbf{R} = \widetilde{\mathbf{r}}]$, настала употребом $\int f_{\boldsymbol{\theta}}(y \mid \mathbf{x} = \widetilde{\mathbf{x}}) I_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}), \widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y}$ (условне расподеле података), не зависи од вредности у $\widetilde{\mathbf{x}}$ које недостају.

Уколико бисмо на свим референтним местима обичну густину/закон расподеле података заменили условним, и теорема 2.3 наставила би да важи. Штавише, не мора чак ни да важи RMCAR, већ слабији услов:

$$\big(\forall \boldsymbol{\phi}\big)\big(\forall \mathbf{y}, \mathbf{y}^* : h(\mathbf{y}) = h(\mathbf{y}^*) = \widetilde{\mathbf{x}}\big) \;\; g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) = g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}^*).$$

2.4 Неопходни услови за игнорабилност

Уочимо да су нам теореме 2.1 и теорема 2.2 дале само **довољне** услове за игнорабилност механизма недостајања при закључивању директно на основу веродостојности и бајесовском закључивању. Теорема 2.1 се, рецимо, бави довољним условима при којима је $\mathcal{L}_{3,\boldsymbol{\phi}}(\boldsymbol{\theta})$ пропорционално са $\mathcal{L}_2(\boldsymbol{\theta})$. Подсетимо се, механизам недостајања моделован помоћу $g_{\boldsymbol{\phi}}(\mathbf{y} \mid \widetilde{\mathbf{r}})$ је само хипотетички, претпостављени модел, индексиран параметрима $\boldsymbol{\phi}$. Стога није неразумно претпоставити да постоји неки ужи, рестрикованији скуп вредности $\boldsymbol{\phi}$, такав да му „право" $\boldsymbol{\phi}$ (у смислу да за то $\boldsymbol{\phi}$ механизам недостајања одговара стварности) припада, а да, можда, RMAR претпоставка не важи. Из претходно реченог видимо да би било згодно имати, макар за неке видове закључивања, карактеризацију игнорабилности, то јест неопходне и довољне услове под којима она важи.

Претпоставимо, само на тренутак, да је $\mathbf{Y}$ дискретно. Тада можемо писати да је

$$\begin{aligned}\mathbf{P}_{\boldsymbol{\theta}}\{\mathbf{Y} = \mathbf{y}\} &= \mathbf{P}_{\boldsymbol{\theta}}\big\{\overline{o}(\mathbf{Y}, \mathbf{R}) = \overline{o}(\mathbf{y}, \mathbf{r}), o(\mathbf{Y}, \mathbf{R}) = o(\mathbf{y}, \mathbf{r})\big\} \\ &= \mathbf{P}_{\boldsymbol{\theta}}\big\{o(\mathbf{Y}, \mathbf{R}) = o(\mathbf{y}, \mathbf{r})\big\}\mathbf{P}_{\boldsymbol{\theta}}\big\{\overline{o}(\mathbf{Y}, \mathbf{R}) = \overline{o}(\mathbf{y}, \mathbf{r}) \mid o(\mathbf{Y}, \mathbf{R}) = o(\mathbf{y}, \mathbf{r})\big\}.\end{aligned}$$

Отклонимо ли се од ове привремене претпоставке, горњу једнакост ћемо лако прилагодити да поприми облик

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{o(\mathbf{Y},\mathbf{R})}\left(o(\mathbf{y}, \mathbf{r})\right) f_{[\overline{o}(\mathbf{Y},\mathbf{R}) \mid o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y}, \mathbf{r}) \mid o(\mathbf{y}, \mathbf{r})\right). \tag{2.8}$$

Сетимо ли се дефиниције за $\mathcal{L}_1$ и уврстимо $\mathbf{r} = \widetilde{\mathbf{r}}$, добијамо

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \int f_{\boldsymbol{\theta}}(\mathbf{y}) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\ &= \int f_{o(\mathbf{Y},\mathbf{R})}\left(o(\mathbf{y}, \widetilde{\mathbf{r}})\right) f_{[\overline{o}(\mathbf{Y},\mathbf{R}) \mid o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\mathbf{y}, \widetilde{\mathbf{r}})\right) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\ &= f_{o(\mathbf{Y},\mathbf{R})}\left(o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\right) \int f_{[\overline{o}(\mathbf{Y},\mathbf{R}) \mid o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\mathbf{y}, \widetilde{\mathbf{r}})\right) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y}.\end{aligned}$$

Слично, $\mathcal{L}_2(\boldsymbol{\theta})$ може се записати као

$$\mathcal{L}_2(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}(\mathbf{y}) I_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\}}(\mathbf{y}) \, \mathrm{d}\mathbf{y}$$

$$
\begin{aligned}
&= \int f_{o(\mathbf{Y},\mathbf{R})}\left(o(\mathbf{y},\widetilde{\mathbf{r}})\right) f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
&= f_{o(\mathbf{Y},\mathbf{R})}\left(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\right) \int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
&= f_{o(\mathbf{Y},\mathbf{R})}\left(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\right),
\end{aligned}
$$

јер се густина проинтеграли на јединицу. Да бисмо дали неопходне и довољне услове за игнорабилност, присетимо се дефиниције комплетности фамилије расподела.

Дефиниција 2.6. Нека је $f_{\boldsymbol{\lambda}}(\cdot)$, $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ фамилија густина или закона расподеле. Кажемо да је она комплетна, уколико важи импликација

$$
(\forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda})(\forall \text{ Борелову } h) \int h(\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y})\mathrm{d}\mathbf{y} = 0 \quad \Longrightarrow \quad (\forall \mathbf{y})\, h(\mathbf{y}) = 0.
$$

Напомена. Уочимо да за фамилију кажемо да је комплетна ако је, заправо, једина непристрасна оцена нуле - сама нула.

Сада ћемо изложити теорему која даје карактеризацију игнорабилности у извесном сценарију. На њој смо захвални ауторима из [15] и [10].

Теорема 2.4. ***Претпоставимо да је*** $\boldsymbol{\Omega}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \boldsymbol{\Omega}_{\boldsymbol{\theta}} \times \boldsymbol{\Omega}_{\boldsymbol{\phi}}$ ***и да*** $f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}}))$ ***представља комплетну фамилију. Нека је још и*** $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \widetilde{\mathbf{y}}) > 0$ ***за све*** $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$***. Тада су, за свако*** $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$***,*** $\mathcal{L}_1(\boldsymbol{\theta},\boldsymbol{\phi})$ ***и*** $\mathcal{L}_2(\boldsymbol{\theta})$ ***пропорционални ако и само ако је недостајање RMAR.***

Доказ. Обрнут смер следи одмах из теореме 2.1. Докажимо директан смер. Нека су за свако $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$, $\mathcal{L}_1(\boldsymbol{\theta},\boldsymbol{\phi})$ и $\mathcal{L}_2(\boldsymbol{\theta})$ пропорционални. Онда мора бити да за свако $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$ важи да

$$
\boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\boldsymbol{\phi}) := \int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y}
$$

не зависи од $\boldsymbol{\theta}$. По дефиницији је

$$
\int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} - \boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\boldsymbol{\phi}) = 0,
$$

а ништа се не мења ако умањилац помножимо јединицом (интегралом густине):

$$
\begin{aligned}
\int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
- \boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\boldsymbol{\phi}) \int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} = 0.
\end{aligned}
$$

Након што мало средимо, добијамо:

$$
\int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right) \left[g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) - \boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\boldsymbol{\phi})\right] I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} = 0.
$$

Приметимо још и да је свеједно да ли под интегралом стоји $f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right)$, или пак $f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}) \mid o(\mathbf{y},\widetilde{\mathbf{r}})\right)$, због индикатора који постоји као чинилац, и свакако остављa само оне $\mathbf{y}$ који се поклапају са $\widetilde{\mathbf{y}}$ на доступним позицијама (оним које нису недостајуће). Ово нам коначно даје да је

$$
\int f_{[\overline{o}(\mathbf{Y},\mathbf{R})|o(\mathbf{Y},\mathbf{R})]}\left(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\right) \left[g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) - \boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\boldsymbol{\phi})\right] I_{\{\mathbf{y}|o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}}(\mathbf{y})\,\mathrm{d}\mathbf{y} = 0,
$$

па се можемо позвати на претпоставку комплетности и закључити да је

$$g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y}) = \boldsymbol{\tau}(\widetilde{\mathbf{r}}, o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \boldsymbol{\phi})$$

за све $\boldsymbol{\phi} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}}$ и оне $\mathbf{y}$ који се са $\widetilde{\mathbf{y}}$ поклапају на доступним пољима. Одавде закључујемо да $g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid \mathbf{y})$, будући исто за разне „недостајуће" делове од $\mathbf{y}$ при фиксном доступном делу (истом као у $\widetilde{\mathbf{y}}$), никако не може зависити од $\overline{o}(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})$, па по дефиницији мора бити да важи RMAR. ■

Напомена (дискусија). Из горње теореме видимо да су за закључивање директно на основу веродостојности, при подацима који имају комплетну расподелу, неопходни и довољни услови за игнорабилност механизма недостајања - RMAR и раздвојеност параметара. Да би игнорабилност важила и за фреквенционистичко закључивање на основу веродостојности, пропорционалност која је поменута мора важити при **свакој** реализацији $\mathbf{y}$, па је за тај вид закључивања EMAR неопходан услов.

Напомена. Постојање слабијих услова од априорне независности параметара за игнорабилност у бајесовском закључивању и даље је отворено питање.

Глава 3

Једнострука импутација

3.1 Откуд потреба за импутацијом?

У претходној глави смо видели, поред игнорабилности, и како се приступом модела решава проблем недостајућих података. Заиста, при таквом приступу дошло се чак и до теорема које тврде карактеризационе услове за игнорабилност механизма недостајања. Стога је потпуно разумно поставити питање зашто бисмо уопште користили неке друге приступе.

У животу, ђаво је у детаљима, а у математици - у практичној примени. Наиме, проблем је у томе што ми у великом броју случајева уопште и **не знамо расподелу података**, већ имамо само податке као вектор или матрицу. У таквим ситуацијама методи засновани на моделима потпуно су немоћни. Ипак, из теорије узорака познато нам је да се узорковање може посматрати на два различита начина[1]:

1. Извучени узорак $\mathbf{y}$ представља реализацију случајне променљиве $\mathbf{Y}$ која има расподелу $f_{\boldsymbol{\theta}}(\cdot)$ која је потенцијално непозната;

2. Постоји коначна популација јединки, обима N из које се извлаче узорци обима n (са или без понављања, стратификовани итд); подаци које анализирамо су један од тих узорака. Тада је сваком узорку додељена вероватноћа извлачења, и једина расподела која нас занима је управо та - уопште није неопходно знати (нити чак дефинисати као концепт) расподелу самих података, већ се говори о тзв. *плану узорковања*. Ни он, разуме се, не мора бити познат.

Било би, дакле, пожељно, имати метод санације недостајања у подацима који је мало или никако зависан од њихове расподеле. Импутација је једно од решења, чак и најчешће коришћено, јер се може задати чак и *алгоритмом*, без потребе за спецификацијом било каквих модела. О овоме ћемо много више рећи у наставку.

Напомена. Уочимо још једну разлику приступа модела и импутације: у првом смо се максимално трудили да недостајуће податке занемаримо, игноришемо, а код импутације им правимо модел и трудимо се да их што верније реконструишемо.

3.2 Преглед фундаменталних резултата у области импутације

Подсетимо се, на самом почетку овог текста кратко смо се осврнули на најчешће коришћен приступ у раду са недостајућим подацима, а то је била *импутација*, или *попуњавање* истих. Термин импутација у српски језик дошао је директно од енглеског термина *imputation* и одомаћио се, док природнији термин *попуњавање* из неког разлога није заживео. Ми ћемо користити оба, како бисмо избегли често понављање исте речи и тиме олакшали читљивост.[2]

[1] Одговарајући термини на енглеском језику су *model-based approach* и *design-based approach*.

[2] Такво понављање у лингвистици је познато као *таутологија*. Не треба га мешати са истоименим математичким појмом.

Импутација, дакле, представља попуњавање недостајућих поља у подацима заменским вредностима. Одмах се намеће много питања: како вршити такво попуњавање, како мерити његов квалитет, како ће оно утицати на валидност статистичког закључивања итд.

За родоначелника импутације као гране статистике и математике сматра се Доналд Б. Рубин (1947-), професор емеритус Универзитета Харвард, који је поставио темеље проучавању недостајућих података у свом раду [31] и књизи [28]. Заједно са својим коаутором Родериком Ј.А. Литлом (1949-) дао је дефиниције MCAR, MAR и MNAR недостајања које смо видели у ранијим одељцима[3], а круну њихове сарадње представља књига [5]. Извршили су значајан утицај на Стефа ван Бурена, који је аутор књиге [6] у којој је, за разлику од [5] акценат посвећен искључиво импутацији као начину рада са недостајућим подацима. Такође, Бурен је у поменутој књизи дао детаљне описе мноштва коришћених алгоритама, а лично је, заједно са Карин Гротиус-Оудсхорн аутор `mice` пакета за програмски језик `R`, што представља најчешће коришћени алат у раду са недостајућим подацима ([14]).

Доприноси импутацији који су побројани у горњем параграфу јесу најбитнији, али свакако нису једини, а ван оквира овог текста излази да се они наброје сви. Знатижељни се свакако упућују на [5] и [6], где се може наћи скоро па свеобухватан преглед литературе на тему импутације, као и рада са недостајућим подацима, закључно са 2019. годином.

3.3 Шта је једноструко у једнострукој импутацији?

Нека имамо податке, као и раније, дате вектором/матрицом $[y_{ij}]$ и нека за неке i и j вредност y_{ij} недостаје. Када кажемо да смо једноструко импутирали недостајућу вредност, подразумевамо да смо, на основу неког модела или алгоритма који смо формирали, на упражњено место уписали заменску вредност. Чему онда потреба за квалификацијом *једноструко*? Та потреба настаје из чињенице да постоји и *вишеструка импутација*, код које се упражњено поље попуњава коришћењем више различитих вредности, то јест врши се m једноструких импутација и добија m комплетираних скупова података, над којима се онда врше класичне статистичке анализе. Вишеструкој импутацији посветићемо наредну главу.

Предности једноструке импутације јесу у томе што су сами алгоритми углавном једноставнији, добија се само један скуп података, па немамо потешкоће са накнадним комбиновањем оцена итд, али такође могу доћи са озбиљним манама, које се углавном огледају у томе што су саме оцене параметара неквалитетне, а изведени закључци непоуздани. Ово ћемо мало касније размотрити и на конкретним примерима.

3.4 Методи једноструке импутације

До сада смо само описали да вршимо некакво попуњавање недостајућих података заменским, али нисмо се осврталиа на то како да то урадимо. Дошао је моменат да скинемо вео мистерије и са тога. Импутација треба бити схваћена као *извлачење узорка из предиктивне расподеле недостајућих вредности*. Дакле, ми ћемо свакој недостајућој вредности приписати расподелу вероватноћа, надамо се тако да она што боље ослика саму расподелу праве вредности, те ћемо вући вредност из те расподеле и њоме импутирати.

Процес формирања такве расподеле, коју смо назвали *предиктивна расподела*, није нимало једноставан. Литл и Рубин ([5]) све начине за генерисање ове расподеле угрубо сврставају у две категорије:

- **Експлицитно моделовање.** Предиктивна расподела је експлицитно задата (нормална, експоненцијална, условна у односу на доступне податке итд.) и из ње се узоркују импутационе вредности. Касније ћемо кроз неколико често коришћених примера појаснити како се задаје ова расподела, како бисмо илустровали шта све може поћи по злу ако се импутација обави на погрешан начин.

[3] Истини за вољу, Рубин оригинално није разликовао RM(C)AR и EM(C)AR, што је у каснијој литератури знало да створи забуну, јер се MAR користило и за EMAR и за RMAR. Ипак, временом су се ствари искристалисале, и данас до забуне углавном не долази. Велики допринос овоме дао је рад [10] који је послужио као извесна кратка систематизација дотадашњег (2013) знања у области.

- **Имплицитно моделовање.** Задаје се само *алгоритам* који даје импутационе вредности, а сама расподела се не задаје директно, већ је она латентна и последица је самог алгоритма. И за ово ћемо дати неколицину примера.

Наравно, за конкретне потребе развијено је много различитих модела и алгоритама за импутацију, као и различитих унапређења постојећих. Многе од њих није могуће грубо сврстати у једну од горње две категорије, јер имају и једних и других особина. Горњу поделу, стога, треба схватити у најлабавијем могућем смислу. Доста о таквим моделима се може прочитати у [6]. Такође, у последњих пар година пробој у импутацији полако су почеле да праве и неуронске мреже, али се још увек не назире унифицирана теорија која би тај приступ објединила. Неуронске мреже у својој природи јесу регресиони и класификациони модели, па се може рећи да њихова употреба у импутацији више нагиње ка експлицитном моделовању (неуронске мреже су, суштински, параметарски модели).

3.4.1 Примери експлицитног моделовања

Пример 3.1 (импутација средњом вредношћу)**.** Једноставан начин за импутацију недостајуће вредности y_{ij} јесте да се она замени просеком $\overline{y}^{(j)}$ свих доступних вредности у j-тој колони, чији ћемо број означити са $n^{(j)}$. Јасно је да просек попуњене колоне остаје $\overline{y}^{(j)}$. Такође, можемо срачунати да је узорачка дисперзија тако импутиране колоне једнака:

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\overline{y}^{(j)})^2 &= \frac{1}{n-1}\sum_{i:\,m_{ij}=0}(y_{ij}-\overline{y}^{(j)})^2 + \frac{1}{n-1}\sum_{i:\,m_{ij}=1}(y_{ij}-\overline{y}^{(j)})^2 \\
&= \frac{1}{n-1}\sum_{i:\,m_{ij}=0}(y_{ij}-\overline{y}^{(j)})^2 + \frac{1}{n-1}\sum_{i:\,m_{ij}=1}(\overline{y}^{(j)}-\overline{y}^{(j)})^2 \\
&= \frac{n^{(j)}-1}{n-1}\frac{1}{n^{(j)}-1}\sum_{i:\,m_{ij}=0}(y_{ij}-\overline{y}^{(j)})^2 \\
&= \frac{n^{(j)}-1}{n-1}(s^{(j)})^2,
\end{aligned}$$

где је $(s^{(j)})^2$ узорачка дисперзија „потколоне" састављене само од доступних поља. Под RMCAR претпоставком, јасно је да је $(s^{(j)})^2$ постојана оцена праве дисперзије од Y_j, јер је таква и узорачка дисперзија. Видимо да у том случају (RMCAR) безусловна импутација средњом вредношћу потцењује праву дисперзију обележја и то за фактор $\frac{n^{(j)}-1}{n-1}$. Ово је једноставна и добра илустрација како неправилан начин импутације може довести до погрешних статистичких закључака. Сличан утицај овај метод врши и на коваријацију две колоне, о чему се може прочитати у [5].

Пример 3.2 (импутација условном средњом вредношћу за једнодимензионо недостајање)**.** Уколико се присетимо да линеарна регресија моделује условно очекивање $\mathbf{E}(Y|X)$ циљне променљиве у односу на атрибуте, можемо рећи да (линеарно)регресиона импутација заправо представља импутацију *условном* средњом вредношћу, а онда импутацију из претходног примера можемо звати импутација *безусловном* средњом вредношћу.

Претпоставимо да су колоне $Y_1,\ldots,Y_{K-1}$ комплетне, а да недостајући подаци постоје само у последњој колони Y_K. Такође, због претпоставке независности инстанци, можемо се договорити да су првих r врста комплетне, а да у последњих $n-r$ недостаје последње поље. Тада се на основу комплетних врста формира линеарна регресија $Y_K \sim (Y_1,\ldots,Y_{K-1})$ која ће дати коефицијенте $(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_{K-1})$, на основу којих се процењује недостајуће поље y_{iK} као

$$\hat{y}_{iK} = \hat{\beta}_0 + \sum_{j=1}^{K-1}\hat{\beta}_j y_{ij}.$$

Наравно, уколико је нека од променљивих у регресији категоричка врши се одговарајуће кодирање (*dummy*, *one hot* итд) и проблем се своди на овај који смо дискутовали.

Није потребно посебно објашњавати зашто овакви методи који у неком смислу врше „најбоља предвиђања" нису адекватни за статистичко закључивање било ког типа. Као прво, они ће значајно да потцене варијабилност саме циљне колоне, а и појачаће линеарну везу у подацима, која ће се касније одразити на конструкцију било каквих модела на подацима. Нпр. у некој каснијој линеарној регресији лажно ће порасти коефицијенти корелације међу колонама, што ће се после одразити на значајност коефицијената итд.

Пример 3.3 (стохастичка регресиона импутација)**.** Овај метод представља природно уопштење претходног које служи да надомести његове главне недостатке: смањење варијабилности и повећање линеарне зависности међу подацима. Претпоставке су исте као и малопре, а импутација се врши по принципу

$$\hat{y}_{iK} = \hat{\beta}_0 + \sum_{j=1}^{K-1} \hat{\beta}_j y_{ij} + z_{iK},$$

где величина z_{iK} има улогу *шума*, а углавном се вуче из нормалне расподеле са очекивањем 0 и дисперзијом која је једнака резидуалној дисперзији модела који је служио за импутацију. Наравно, постоје и модификације које случајно узоркују један од резидуала и одмах га додају као шум и сл. Сада ћемо навести неколико корисних напомена у вези са стохастичком регресионом импутацијом.

Напомена. Претпоставка о недостајућим подацима у само једној колони је, разуме се, веома ограничавајућа. Срећом постоје разне модификације (стохастичке) регресионе импутације које омогућавају њену примену и у таквим ситуацијама. Суштински говорећи, већина њих се своди на вишеструку примену стохастичке регресионе импутације за различите регресионе коефицијенте и комплетне базе. Једна од таквих модификација јесте Баков метод, о којем се може прочитати у [33]. Ми се нећемо оптерећивати таквим модификацијама да се не бисмо расплинули, али нам је битно да знамо да постоје и да је могуће задржати непристрасност/постојаност оцена.

Напомена. Рубин и Литл ([5]) су показали да је за дводимензиону нормалну расподелу, са недостајањима у другој колони Y_2 удела $\lambda = \frac{n-r}{n}$ и MCAR механизам преферабилно користити стохастичку регресиону импутацију, јер она једина даје постојане оцене и дисперзије Y_2, и $\mu_2 = \mathbf{E}Y_2$, као и регресионих параметара у регресијама $Y_1 \sim Y_2$ и $Y_2 \sim Y_1$ (у поређењу са импутацијом безусловном средњом вредношћу и обичном регресионом импутацијом). У наведеној референци може се наћи и проређење са методима из примера 3.1 и 3.2, где су експлицитно изведене и пристрасности оцена. Ипак, стохастичка регресиона импутација има и две озбиљне мане. Прва се огледа у томе што додавање случајног шума значајно повећава дисперзије оцена, што су Рубин и Литл такође експлицитно извели. Друга мана је та што су стандардне грешке оцена параметара на основу података овим методом премале, јер не урачунавају неодређеност коју доноси импутација. Овај проблем решиће вишеструка импутација којом ћемо се бавити у наредној глави.

Напомена. Видели смо да се претпоставка о једнодимензионом недостајању може пренебрегнути одговарајућом модификацијом стохастичке регресионе импутације. Међутим, то не мора бити случај за претпоставку о RMCAR недостајању. Наиме, проблем настаје у томе што подузорак комплетних опсервација не мора бити репрезентативан, па саме оцене регресионих коефицијената на основу којих се врши импутација постају веома пристрасне. То се у контексту импутације донекле, барем емпиријски, може решити тако што се за сваку појединачну импутацију коефицијенти рачунају поново, али не на основу подузорка комплетних опсервација, већ на основу неког бутстреп узорка извученог из њега. Алтернатива томе било би да се коефицијенти посматрају као случајне величине, па да се при свакој импутацији вредности параметара извлаче из њихових апостериорних расподела, при услову доступних података. Детаљан опис оваквих алгоритама може се пронаћи у [6].

Импутација коришћењем сингуларне декомпозиције

Овај пример биће мало дужи, па заслужује посебан сегмент текста. Овај метод преузели смо из [20] где је он и уведен.

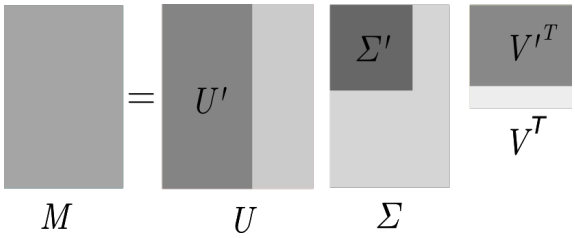
Прво ћемо се подсетити појма сингуларне декомпозиције матрице[4]. Наредна теорема класичан је резултат из Линеарне алгебре и наводимо је без доказа. Доказ се може пронаћи у [22].

Теорема 3.1. ***Нека је*** $\mathbf{M} \in \mathbb{C}^{m\times n}$***. Тада се*** $\mathbf{M}$ ***може представити у облику***

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*,$$

где је $\mathbf{U} \in \mathbb{C}^{m\times m}$ ***унитарна матрица,*** $\mathbf{\Sigma} \in \mathbb{C}^{m\times n}$ ***правоугаона дијагонална матрица са ненегативним елементима на дијагонали, а*** $\mathbf{V} \in \mathbb{C}^{n\times n}$ ***такође унитарна матрица. Уколико је*** $\mathbf{M}$ ***реална матрица, такве су и*** $\mathbf{U}$ ***и*** $\mathbf{V}$ ***(ортогоналне). Такође, увек постоји сингуларна декомпозиција таква да су дијагонални елементи матрице*** $\mathbf{\Sigma}$ ***сортирани нерастуће.***[5] ■

Приметимо да нисмо помињали однос бројева m и n. Пошто је нама у статистици од главне важности дизајн матрица, најчешћи случај ће бити већи број врста од броја колона. Али, чак ни у статистици то не мора увек бити случај (мали обими узорка). Стога, у општем случају матрица $\mathbf{\Sigma}$ потенцијално има $m - r$ нула-врста и $n - r$ нула-колона.



Слика 3.1: Општи случај сингуларне декомпозиције

На слици 3.1 графички је приказан општи случај сингуларне декомпозиције. Матрицу $\mathbf{U}'$ сачињава првих $r = R(\mathbf{M}) = R(\mathbf{\Sigma})$ колона матрице $\mathbf{U}$, горњи леви блок матрице $\mathbf{\Sigma}$ јесте матрица $\mathbf{\Sigma}'$ која је дијагонална и на њеној дијагонали налазе се нерастуће сортиране све ненула сингуларне вредности, њих r. Остатак матрице Σ сачињавају нуле. Врсте матрице $(\mathbf{V}')^T$ представљају првих r колона матрице $\mathbf{V}$. Изостављање било којих врста/колона у матрици $\mathbf{\Sigma}$ повлачи изостављање неких колона у матрицама $\mathbf{U}$ и $\mathbf{V}$. Стога се сингуларна декомпозиција из прве теореме назива још и *пуна* (енг. *full SVD*). Из претходно испричаног намећу се још три типа, а то су такозване *редуковане* сингуларне декомпозиције:

1. Уколико из матрице $\mathbf{\Sigma}$, ако има више врста, избацимо онолико нула-врста колико је вишак да би она била квадратна, а одговарајуће нула-колоне ако има више колона, онда од матрице $\mathbf{\Sigma}$ „остаје" квадратна матрица $\mathbf{\Sigma}_{\min\{m,n\}}$ која је квадратна, димензије $\min\{m, n\}$. Од матрица $\mathbf{U}$ и $\mathbf{V}$ се онда задржава $\min\{m, n\}$ колона (од једне све, а од друге не).

   Оваква декомпозиција зове се *танка сингуларна декомпозиција* (енг. *thin SVD*).

2. Од матрице $\mathbf{\Sigma}$ задржава се само горњи леви дијагонални блок са ненула дијагоналним елементима - матрица $\mathbf{\Sigma}'$. Тада се из матрице $\mathbf{U}$ одбацује десних $m - r$ колона, а из матрице $\mathbf{V}$ десних $n - r$ колона (може се, наравно, десити да се не избаци ниједна).

   Оваква декомпозиција зове се *компактна сингуларна декомпозиција* (енг. *compact SVD*).

3. Од матрице $\mathbf{\Sigma}$ задржава се само горњи леви дијагонални блок $\mathbf{\Sigma}_t$ са ненула елементима, али дужине $t \leqslant r$. Тада се из матрице $\mathbf{U}$ одбацује десних $m - t$ колона, а из матрице $\mathbf{V}$ десних $n - t$ колона. Од матрице $\mathbf{U}$ остаје $\mathbf{U}_t \in \mathbb{R}^{m\times t}$, а од матрице $\mathbf{V}$ остаје $\mathbf{V}_t \in \mathbb{R}^{n\times t}$. Декомпозиција поприма облик

$$\hat{\mathbf{M}} = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^T,$$

[4]енг. *Singular Value Decomposition - SVD*

[5]Што ћемо убудуће и подразумевати. Те дијагоналне елементе зваћемо *сингуларне вредности*.

где наравно $\mathbf{M} \neq \widehat{\mathbf{M}}$ у општем случају. Ипак испоставља се да је матрица $\widehat{\mathbf{M}}$ у извесном смислу „блиска" матрици $\mathbf{M}$, о чему ћемо говорити у даљем тексту.

Оваква декомпозиција зове се *крња сингуларна декомпозиција* (енг. *truncated SVD*).

Напомена. Надаље ћемо претпоставити центрираност података како бисмо избегли слободан члан у неким моделима.

Нека је $\mathbf{Y} \in \mathbb{R}^{n\times p}$ матрица у којој су смештени подаци. Нека је $\mathbf{Y}^c$ матрица која се састоји из свих оних **врста** матрице $\mathbf{Y}$ којима не недостаје ниједно поље и нека је $\mathbf{Y}^m$ матрица која се састоји из оних врста матрице $\mathbf{Y}$ којима недостаје барем једно поље. Нека је дата крња сингуларна декомпозиција матрице $\mathbf{Y}^c$:

$$\widehat{\mathbf{Y}}^c_J = \mathbf{U}_J\mathbf{D}_J\mathbf{V}_J^T,$$

где је $\mathbf{D}_J$ дијагонална матрица која на дијагонали има водећих $J \leqslant p$ ненула сингуларних вредности матрице $\mathbf{Y}^c$, а $\mathbf{U}_J$ и $\mathbf{V}_J$ су одговарајуће матрице са међусобно ортогоналним колонама које садрже првих J колона матрица $\mathbf{U}$ и $\mathbf{V}$ за $\mathbf{Y}^c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

Испоставља се да је матрица $\widehat{\mathbf{Y}}^c$ од велике важности, што ћемо формулисати у следећој теореми, која је такође резултат Линеарне алгебре и нећемо је доказивати.

Теорема 3.2. ***Уз до сада коришћену нотацију, матрица*** $\mathbf{M}$ ***ранга*** J ***која решава проблем***

$$\arg \min_{\substack{\mathbf{M}\\ R(\mathbf{M})=J}} \|\mathbf{Y}^c - \mathbf{M}\|,$$

где $\|\cdot\|$ ***представља*** L^2 ***норму***[6]***, јесте матрица*** $\widehat{\mathbf{Y}}^c$***. Дакле,*** $\widehat{\mathbf{Y}}^c$ ***је она од матрица ранга*** J ***која је најближа матрици*** $\mathbf{Y}^c$ ***у средњеквадратном смислу.*** ■

Напомена. Није битно да ли минимизујемо норму или квадрат норме, јер је квадрат строго растућа функција.

Сада ћемо погледати овај резултат са статистичке тачке гледишта. Нека је y било која врста матрице $\mathbf{Y}^c$, посматрана као колона вектор. Посматрајмо линеарни регресиони модел у којему је y зависна променљива, а $\mathbf{V}_J$ дизајн матрица, односно тражимо

$$\arg\min_{\boldsymbol{\beta}} \|y - \mathbf{V}_J\boldsymbol{\beta}\|.$$

Са курса нам је позната чињеница да је тада одговарајућа оцена дата изразом $\hat{\boldsymbol{\beta}} = (\mathbf{V}_J^T\mathbf{V}_J)^{-1}\mathbf{V}_J^T y$, а како су колоне матрице $\mathbf{V}_J$ међусобно ортогоналне добијамо да је оцена методом најмањих квадрата:

$$\hat{\boldsymbol{\beta}} = \mathbf{V}_J^T y.$$

Одавде закључујемо да се у колонама матрице $\mathbf{V}_J^T(\mathbf{Y}^c)^T$ налазе оцене МНК коефицијената за све могуће моделе са дизајн матрицом $\mathbf{V}_J$ и зависном променљивом која пролази кроз све могуће врсте матрице $\mathbf{Y}^c$. Сходно томе, у **колонама** матрице $\mathbf{V}_J\mathbf{V}_J^T(\mathbf{Y}^c)^T$ налазе се оцењене вредности **врста** матрице $\mathbf{Y}^c$ у поменутој регресији. Коначно, у врстама транспоната те матрице, тј. у врстама матрице $\mathbf{Y}^c\mathbf{V}_J\mathbf{V}_J^T$ налазе се оцене врста матрице $\mathbf{Y}^c$.

Када би било $J = r = R(\mathbf{Y}^c)$, могли бисмо урадити следеће:

$$\mathbf{Y}^c\mathbf{V}_J\mathbf{V}_J^T = \mathbf{U}_J\mathbf{D}_J\mathbf{V}_J^T\mathbf{V}_J\mathbf{V}_J^T = \mathbf{U}_J\mathbf{D}_J\mathbf{V}_J^T = \widehat{\mathbf{Y}}^c,$$

па би матрица која је у средњеквадратном смислу најближа матрици $\mathbf{Y}^c$ била и њена линеарнорегресиона оцена МНК. Међутим, ми имамо крњу декомпозицију. Ипак, одлучујемо се да матрицу $\mathbf{Y}^c$ оценимо баш матрицом $\widehat{\mathbf{Y}}^c$, управо јер желимо да искористимо информацију о минимизацији L^2 растојања.

Оваква (макар и приближна) веза даје нам извесну наду да би се и врста из $\mathbf{Y}^m$, којој (сигурно!) недостају одређена поља, могла оценити на сличан начин. Нека је y_l^T нека од врста

[6] Квадратни корен суме квадрата свих поља.

из матрице $\mathbf{Y}^m$. Избацимо из ње недостајућа поља, транспонујмо је, и добићемо колону y^*. Сада ћемо посматрати регресиони проблем

$$\arg\min_{\boldsymbol{\beta}} \|y^* - \mathbf{V}_J^*\boldsymbol{\beta}\|^2,$$

где је $\mathbf{V}_J^*$ матрица која је настала од $\mathbf{V}_J$ избацивањем оних врста које имају исти редни број као недостајуће поље у y^*. Тада је оцена коефицијената МНК дата са $\hat{\boldsymbol{\beta}} = ({\mathbf{V}_J^*}^T\mathbf{V}_J^*)^{-1}{\mathbf{V}_J^*}^T y^*$, а оцењене вредности дате су са $\hat{y}^* = \mathbf{V}_J({\mathbf{V}_J^*}^T\mathbf{V}_J^*)^{-1}{\mathbf{V}_J^*}^T y^*$. Матрица $\mathbf{V}_J^*$ више **нема ортогоналне колоне**, тако да се средњи производ не може скратити. Оцену ћемо стога записати са $\hat{y}^* = \mathbf{V}_J^*\hat{\boldsymbol{\beta}}$. У овом тренутку као једини смислен начин да се оцене недостајућа поља y_{miss} јавља се

$$\hat{y}_{miss} = \mathbf{V}^{(*)}\hat{\boldsymbol{\beta}},$$

где смо са $\mathbf{V}^{(*)}$ означили комплемент матрице $\mathbf{V}_J^*$ у матрици $\mathbf{V}_J$ ($\hat{y}_{miss}$ је колона оцена недостајућих поља у редоследу у коме недостају). *Овом оценом ћемо се овде задовољити.*

Све што смо до сада рекли је тачно. Ипак, постоје два питања на која је остало да се одговори, а то су:

1. Зашто смо бирали баш регресију са дизајн матрицом $\mathbf{V}_J$?

2. Како одабрати адекватно J?

На друго питање нема универзалног одговора, углавном се до њега долази емпиријски, испробавањем разних вредности. Одговор на прво питање је донекле једноставан, уколико примена регресије почетна идеја. Потребно је алгоритам који смо објаснили „прочитати уназад". Желимо да оценимо поље које недостаје у врсти y_l. Оценићемо целу врсту. Међутим, не можемо целу врсту користити као вектор зависних вредности, јер јој фале поља. Избацимо поља која фале, оценимо остатак. Шта узети за дизајн матрицу? Хајде да погледамо шта има смисла за врсту у којој оригинално нема недостајућих података. Уколико за дизајн матрицу узмемо $\mathbf{V}_J$ из крње сингуларне декомпозиције, и ако је баш $J = r$, добићемо МНК оцену која је у средњеквадратном смислу најближа оригиналној матрици $\mathbf{Y}^c$. Уколико из практичних разлога рачунања колона у $\mathbf{V}_J$ смањимо J оцена више неће бити МНК (иако је близу!) али и даље има својство средњеквадратне блискости. *Алгоритам који смо представили јесте овај параграф, обрнутим редоследом.*

Претходни метод има недостатке који га у многим ситуацијама могу учинити потпуно бескорисним. Наиме, у многим реалним ситуацијама већина врста у матрици предиктора има неко недостајуће поље. Тако нам се може десити да за импутацију користимо свега, рецимо, једну трећину података. Такође, чак и да нам недостаје толико много података, ипак је јасно да исувише тога остаје неискоришћено. Стога ћемо наћи начин како да искористимо и врсте из матрице $\mathbf{Y}^m$. Овај пут ћемо у поставци проблема **изоставити претпоставку центрираности**, јер није могуће центрирати врсту ако јој поље недостаје.

Решавамо следећи проблем:

$$\arg \min_{\substack{\mathbf{M} \\ R(\mathbf{M})=J}} \|\mathbf{Y} - m\cdot\mathbf{1}^T - \mathbf{M}\|_*,$$

где је $\|\cdot\|_*$ норма матрице која представља корен из суме квадрата свих присутних поља (дакле оних која не недостају). $\mathbf{1}$ представља колону јединица, а m вектор средњих вредности врста од $\mathbf{Y}$ (поља могу да недостају, ускоро ћемо видети шта тада). Производ $m\cdot\mathbf{1}^T$ заправо представља матрицу истих димензија као $\mathbf{Y}$, где свака врста има међусобно иста поља, а она одговарају средњој вредности те врсте у $\mathbf{Y}$. Другим речима:

$$m\cdot\mathbf{1}^T = \begin{bmatrix} \bar{y}_{1\rightarrow} \\ \vdots \\ \bar{y}_{N\rightarrow} \end{bmatrix} \cdot \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \bar{y}_{1\rightarrow} & \cdots & \bar{y}_{1\rightarrow} \\ \vdots & \ddots & \vdots \\ \bar{y}_{N\rightarrow} & \cdots & \bar{y}_{N\rightarrow} \end{bmatrix}.$$

За импутацију примењујемо следећи итеративни алгоритам:

1. На почетку недостајуће податке заменимо средњом вредношћу одговарајуће врсте да бисмо добили комплетну матрицу $\mathbf{Y}^0$. Поставимо $i = 0$ (нулти корак).

2. $\mathbf{Y} - m \cdot \mathbf{1}^T$ сада представља по врстама центрирану матрицу[7], тако да $\mathbf{M}$ можемо оценити на начин као раније и (за одабрано J) добити матрицу $\mathbf{Y}^i$ за оцену матрице $\mathbf{Y}^0$ (тако што ћемо на матрицу која је регресиона оцена додати оно што смо одузели: $m \cdot \mathbf{1}^T$). Из ње узимамо она поља која фале у $\mathbf{Y}$ и стављамо их на празна места у $\mathbf{Y}$. Тиме смо добили матрицу $\mathbf{Y}^{i+1}$.

3. Инкрементирамо бројач и настављамо са итерацијом. Итерирамо све док $\frac{\|\mathbf{M}^i - \mathbf{M}^{i+1}\|}{\|\mathbf{M}^i\|}$ не постане мање од неког предефинисаног прага толеранције ε, где нам $\mathbf{M}^i$ представља целу оцењену матрицу у i-том кораку, дакле са слободним чланом (тј. додато јој је оно што је пре регресије одузето, а то је $m \cdot \mathbf{1}^T$).

Напомена. Заиста изгледа чудно и неприродно користити врсту за тренирање линеарне регресије, те сам овај метод делује јако проузвољно. Ипак, он се у [20] уводи специфично у контексту података у којима свака врста представља експресију једног гена. Овај контекст превазилази оквире овог рада и залази у поље биостатистике, те се ми њиме даље нећемо бавити. Ипак смо се одлучили да и овај пример интегришемо у рад, јер сматрамо да неком може послужити као мотивација за бављење биостатистиком.

3.4.2 Примери имплицитног моделовања

Хот дек фамилија метода

Суштина *хот дек* (енг. *hot deck*) импутације огледа се у томе да се недостајућа вредност y_{ij} мења неком вредношћу y_{kj} из података, која је доступна, али таква да су врсте y_i и y_k у извесном смислу блиске. Дакле, за инстанцу y_i која садржи недостајућу вредност y_{ij} бира се тзв. *донорска* инстанца (врста) y_k и њено одговарајуће поље користи се за импутацију. Различити *хот дек* методи разликују се по томе на који начин ће се бирати та донорска инстанца. Ми ћемо навести неке од њих.

Пример 3.4 (*хот дек* за једнодимензиони узорак - оцена просека)**.** Претпоставимо да је свака опсервација дужине 1, то јест да меримо само једно обележје. Претпоставимо да је број врста у нашим подацима једнак n и да је узоркован из популације обима N (дакле *design-based* истраживање). Због једноставности, а не губећи општост, сматраћемо да се у узорку налази првих n јединки из популације. Нека је првих $r < n$ вредности узорка доступно, а остале нека недостају.

При оваквим претпоставкама, средња вредност обележја може се оценити као

$$\overline{y}_{HD} = \frac{r\overline{y}_R + (n-r)\overline{y}^*_{NR}}{n},$$

где је $\overline{y}_R$ средња вредност доступних јединки у узорку[8], док је

$$\overline{y}^*_{NR} = \frac{1}{n-r}\sum_{i=1}^{r} H_i y_i,$$

а H_i представља број пута колико је y_i коришћено као донор. Приметимо битну ствар: не импутира се свака недостајућа вредност вредношћу $\overline{y}^*_{NR}$, већ неком конкретном вредношћу y_i, а $\overline{y}^*_{NR}$ је само згодан начин да се оцена просека запише кратком формулом. Јасно, $\sum_{i=1}^{r} H_i = n - r$.

Особине оцене $\overline{y}_{HD}$ зависиће од тога колико квалитетно смо одабрали бројеве H_i. Како се овде налазимо у „дегенерисаном" случају хот дек импутације, то јест немамо како да поредимо врсте, јер или су целе присутне, или целе одсутне, то је једино што можемо да вредностима

[7] Делује неприродно центрирати по врстама, а не по колонама, али то има смисла кад знамо да ће нам врсте представљати зависну променљиву у линеарној регресији.

[8] R је од енглеског *respondent*, тј. „онај који је одговорио". Слично, NR је скраћено од *non-respondent*. Ознаке су Литлове и Рубинове из [5].

$(H_1, \ldots, H_r)$ припишемо расподелу вероватноћа, условну при доступним подацима (њих гледамо као константе). Означимо доступне податке, као и раније, са $o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})$, где је $\widetilde{\mathbf{y}}$ реализација. Тада се очекивање и дисперзија оцене $\overline{y}_{HD}$ могу записати као:

$$\mathbf{E}\,\overline{y}_{HD} = \mathbf{E}_Y \mathbf{E}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})), \tag{3.1}$$

$$\mathbf{D}\,\overline{y}_{HD} = \mathbf{D}_Y \mathbf{E}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) + \mathbf{E}_Y \mathbf{D}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})). \tag{3.2}$$

Друга једнакост представља познату формулу теорије вероватноћа за декомпозицију дисперзије. Други сабирак у последњој једнакости представља додатак на дисперзију оцене који је последица тога што импутирамо, а немамо праве вредности.

Претпоставимо сада још и да вредности $(H_1, \ldots, H_r)$ бирамо као прост случајан узорак са понављањем. Прецизније, импутационе вредности бирамо као ПСУ са понављањем, а онда су све вредности H_i једнозначно одређене. Може се показати[9] ([30]) да при услову доступних података $(H_1, \ldots, H_r)$ има мултиномијалну расподелу са дужином $n-r$ и свим вероватноћама једнаким $1/r$. Знајући то, лако се рачуна да је

$$\mathbf{E}(H_i \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \frac{n-r}{n},$$

$$\mathbf{D}(H_i \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \frac{(n-r)(1-1/r)}{r},$$

$$\mathbf{Cov}(H_i, H_j \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = -\frac{n-r}{r^2}.$$

Када знамо ово, оцену $\overline{y}_{HD}$ „нападнемо" условним очекивањем $\mathbf{E}(\cdot \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}))$, и срачунамо да је

$$\mathbf{E}(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \overline{y}_R$$

и

$$\mathbf{D}(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \left(1 - \frac{1}{r}\right)\left(1 - \frac{r}{n}\right)\frac{s_R^2}{n},$$

где је s_R^2 узорачка дисперзија на доступним инстанцама. Следећи корак јесте претходне две срачунате величине „напасти" очекивањем у односу на расподелу узорка. Ту ситуација може поћи по злу, у зависности од тога који је механизам недостајања и начин на који је сам узорак који представља наше податке извучен. Рецимо, кад недостајање не би било RMCAR, $\overline{y}_R$ врло вероватно не би било непристрасна и постојана оцена просека. Али, ако наметнемо претпоставку RMCAR недостајања, као и то да је узорак ПСУ са понављањем, можемо добити да је

$$\mathbf{E}\,\overline{y}_{HD} = \mathbf{E}_Y \mathbf{E}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \overline{y} \tag{3.3}$$

и

$$\mathbf{D}\,\overline{y}_{HD} = \mathbf{D}_Y \mathbf{E}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) + \mathbf{E}_Y \mathbf{D}_H(\overline{y}_{HD} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})) = \left(\frac{1}{r} - \frac{1}{N}\right) S^2 + \left(1 - \frac{1}{r}\right)\left(1 - \frac{r}{n}\right)\frac{S^2}{N}, \tag{3.4}$$

где је $\overline{Y}$ права популацијска средина, а S^2 права популацијска дисперзија.[10]

Из (3.3) видимо да је, при RMCAR и ПСУСП претпоставкама, $\overline{y}_{HD}$ непристрасна оцена и ту се сва дискусија завршава. Оно што је занимљиво за дискусију јесте формула (3.4). Први сабирак не представља ништа друго неголи узорачку дисперзију оцене $\overline{y}_R$ на основу ПСУ са понављањем. Други сабирак, заправо, представља додатак на дисперзију хот дек оцене

[9] Дата је референца, али је мање-више очигледно.

[10] Рачуна се као $\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$, јер смо претпоставили дизајн експеримента у којем је вредност обележја константна, а случајан је узорак који се извлачи.

која је настала као последица извлачења импутација. То смо у општем случају имали дато у (3.2), а при додатним претпоставкама га имамо у много конкретнијем облику и можемо га користити за извођење закључака. Јасно, S^2 не мора бити познато, али може бити оцењено.

Додатак на дисперзију који се појављује може бити умањен одабиром другог начина узорковања, нпр. простог случајног узорковања без понављања. Ми се тиме овде нећемо бавити, јер смо и овим примером фино илустровали како овај метод утиче на статистичко закључивање у случају оцене просека и њене дисперзије. Више о томе може се наћи у [5].

Напомена. Није тешко уочити да се овакав метод може применити и на податке који имају више атрибута, тако што се посматра колона од интереса. Међутим, јасно је да се ту губи доста информације, јер би остали атрибути потенцијално могли бити инкорпорирани у оцену, како би побољшали њен квалитет. Штавише, показано је да су хот дек оцене које смо поменули у примеру 3.4 непристрасне само за RMCAR недостајање, што је често нереалистична претпоставка. Због тога су развијена унапређења, од којих једно дајемо у наредном примеру.

Пример 3.5 (*хот дек* заснован на растојању)**.** Претпоставимо да имамо вишедимензионе податке, али да недостајања имамо само у једној колони. Ту једну колону зваћемо Y, а остале $X_1, \ldots, X_K$. У овом случају стратегија ће нам бити да дефинишемо метрику $d(x_i, x_j)$ која ће да мери растојање између инстанци $x_i = (x_{i1}, \ldots, x_{ik})$ и $x_j = (x_{j1}, \ldots, x_{jk})$. Уколико су сви атрибути нумерички онда можемо бирати еуклидску метрику, а иначе се могу правити разне модификације. Уколико дефинишемо праг толеранције d_0, недостајућу вредност y_i импутираћемо неком од вредности y_j, тако да буде испуњено $d(x_i, x_j) < d_0$. На овај начин можемо добити више потенцијалних инстанци као кандидате, па се на разне начине можемо одлучити како да одаберемо једну (да случајно узоркујемо, да смањимо d_0 док не остане само једна итд).

Разни избори растојања постоје, а ми ћемо у наредном примеру описати један занимљив избор.

Пример 3.6 (*Predictive Mean Matching*)**.** Нека су претпоставке као у претходном примеру. Нека је задато растојање[11] међу инстанцама као:

$$d(x_i, x_j) = (\hat{y}(x_i) - \hat{y}(x_j))^2,$$

где је $\hat{y}(x_i)$ линеарнорегресиона оцена од y_i за коју су коефицијенти добијени само на основу комплетних врста (као код регресионе импутације). хот дек импутација са овако задатом метриком на простору инстанци у литератури се назива *Predictive Mean Matching*, а најбољи превод, ако себи дамо слободу, могао би бити *импутација најближом инстанцом у регресионом слислу.* Ми ћемо користити скраћеницу PMM.

Да видимо како алгоритам ради и зашто се овако зове. Прођимо кроз њега корак по корак.

1. Прво се формира линеарни регресиони модел $Y \sim (X_1, \ldots, X_K)$ на основу комплетних врста и одатле се добију регресиони коефицијенти $\hat{\beta}$.

2. Изврши се предикција овим моделом за сваку од инстанци. Другим речима, све инстанце се трасформацијом

$$\hat{y}_j = \beta_0 + \sum_{i=1}^{K} \beta_i x_{ji}$$

   пресликају у скуп $\mathbb{R}$. То се уради и са инстанцом са недостајућом вредношћу, рецимо да је то i-та инстаца.

3. Сада имамо вредност предикције $\hat{y}_i$ за некомплетну инстанцу и вредности предикција за остале инстанце.

[11] Истини за вољу, при оваквој дефиницији може бити да нека од особина метрике не буде задовољена. Рецимо, може се десити да различите инстанце буду исто предвиђене, па да им растојање буде нула. Тиме „пуца" недегенерисаност норме. Међутим, овако дефинисано растојање нама не квари ништа, јер је и даље симетрично и задовољава неједнакост троугла, што су нама главне особине.

4. За доноре се бирају све **оригинално комплетне** инстанце чије су пројекције на растојању d_0 или мањем од оне у коју импутирамо. Од њих се на случајан начин бира једна и њено y (**не предикција**) се употреби за попуњавање.

Овај алгоритам је доста сродан стохастичкој регресионој импутацији, али се ипак сврстава у имплицитне методе ([6]). Постоје три основне предности у односу на стохастичку регресиону импутацију. Прва је та што нема тако уочљиво повећање линеарне зависности међу атрибутима. Друга предност је та што је мање осетљив на погрешан одабир модела. Наиме, код регресионе импутације, уколико линеаран модел који користимо за импутацију заправо није адекватан, неће бити ни импутиране вредности. Овде ипак имамо још један слој заштите, јер импутирамо вредностима које већ постоје у подацима. Трећа предност је та што се неће добити немогуће вредности. На пример, (стохастичка) регресиона импутација често ће предвидети негативне висине, масе, а још чешће уделе или уопште ствари блиске нули које су по својој природи строго позитивне. Основне мане алгоритма су те што су симулације показале да лоше ради за мали број атрибута, а да за мали обим узорка долази до дуплицирања донора. Такође, јасно је да колона у коју се импутира мора бити нумеричка.

Модификације алгоритма су разне. Прво, понекад се не бирају за доноре инстанце које су на растојању d_0 или мање од циљне, већ се бира l најближих. Пракса је показала да је углавном оптимално $l = 5$. Такође, свакој инстанци се може приписати вероватноћа, углавном обрнуто сразмерна растојању од циљне инстанце, па се онда донор узорковати у складу са тако добијеном расподелом вероватноћа. Неретко се коефицијенти $\hat{\beta}$ који су добијени на основу комплетних врста не користе за предикцију на некомплетним, како бисмо се заштитили од нерепрезентативности подузорка комплетних врста (тј. тога да недостајање можда није RMCAR). Неке од алтернатива су да се на некомплетним врстама предикција врши уз помоћ коефицијената $\tilde{\beta}$ који су настали на основу бутстреп узорка комплетних врста. Оно што је имплементирано у најчешће коришћеном рачунарском пакету за импутацију, `mice`, јесте да се коефицијентима β припише априорна расподела, па да се $\tilde{\beta}$ вуче из апостериорне расподеле при услову доступних података. Овакво различито рачунање предвиђања за комплетне и некомплетне инстанце познато је под називом *matching* (упаривање), па отуда и име алгоритма. Више о овим варијантама може се прочитати у [6].

Треба рећи још и да се PMM може посматрати као шира класа алгоритама која има заједничку структуру: изврши се предвиђање на основу неког модела, и еуклидско растојање предвиђања користи се за растојање међу инстанцама. У пракси је, ипак, најчешће коришћен модел линеарна регресија.

Алгоритам није тешко модификовати да ради са недостајањима у више колона. Неке од могућих модификација су:

- Да се комплетне колоне користе као предиктори. Дакле, нека су $(Y_1, \ldots, Y_r)$ некомплетне колоне, а $(X_1, \ldots, X_s)$ колоне које су комплетне. Растојања међу инстанцама рачунају се само на основу колона $(X_1, \ldots, X_s)$.

- Ако све, или скоро па све, колоне имају недостајуће податке, онда се може посматрати посебан хот дек за сваку инстанцу: констатујемо циљну инстанцу у коју хоћемо да импутирамо, а затим све инстанце које имају доступне оне ћелије које има и она; међу њима бирамо доноре на основу неке одабране метрике.

Више о овим прилагођавањима алгоритма може се прочитати у [5].

Да бисмо демонстрирали предност алгоритма *Predictive Mean Matching* у односу на импутацију линеарном регресијом[12], генерисаћемо синтетичке податке. Узорковали[13] смо 50 независних инстанци из дводимензионе[14] нормалне расподеле са вектором очекивања и коваријационом матрицом датим са:

[12]Нећемо поредити са стохастичком регресионом импутацијом, јер за њу важи поента о очувању реалне линеарне повезаности коју желимо да демонстрирамо. Њене су мане друге природе, а поменули смо их раније у примеру.

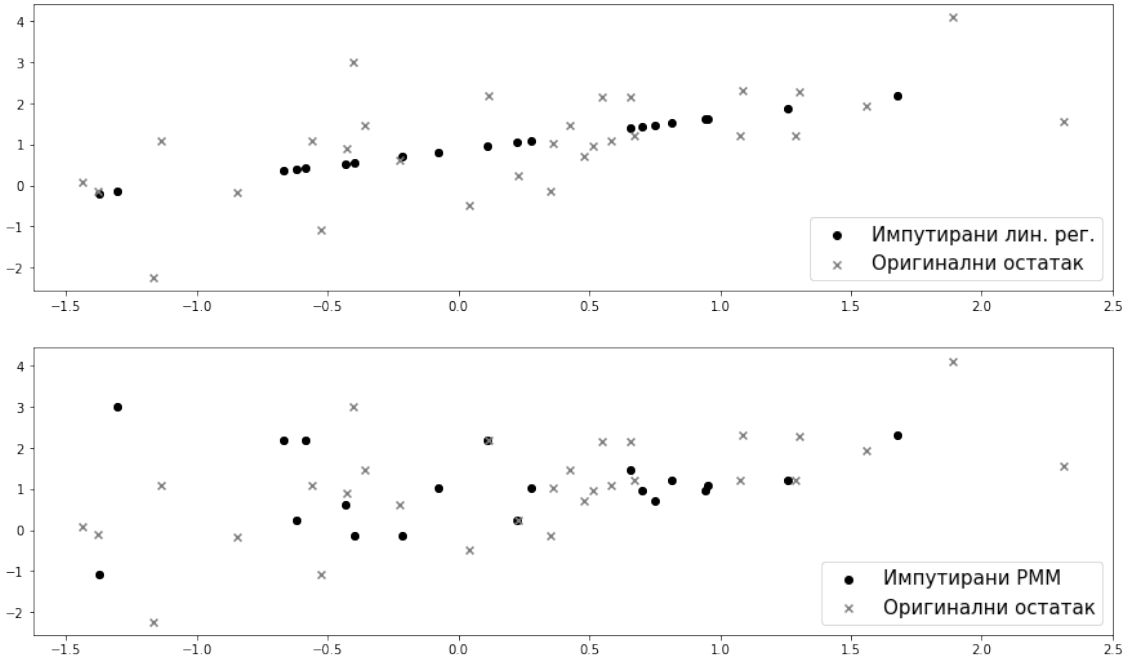
[13]За саму обраду података коришћен је програмски језик `R`, а посебно библиотеке `mvtnorm`, `missMethods` и `mice`, док је за графички приказ коришћена библиотека `matplotlib` програмског језика `Python`.

[14]Да би могло лепо да се види на слици. Поента, наравно, стоји и за друге димензије.

$$\boldsymbol{\mu} = (0, 1) \quad \text{и} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 2 \end{bmatrix},$$

те смо у другој колони генерисали 40% недостајућих података према (R)MCAR механизму.

На слици 3.2 видимо сивим иксевима означене податке који нису имали недостајања, а црним тачкама импутиране инстанце. Горњи график одговара регресионој импутацији, а доњи алгоритму PMM.



Слика 3.2: Поређење линеарно-регресионе и PMM импутације за узорак из дводимензионе нормалне расподеле: импутиране вредности и комплетни остатак

На слици 3.3, за исте податке и исто генерисано недостајање дат је приказ импутираних података и њихових одговарајућих оригинала.

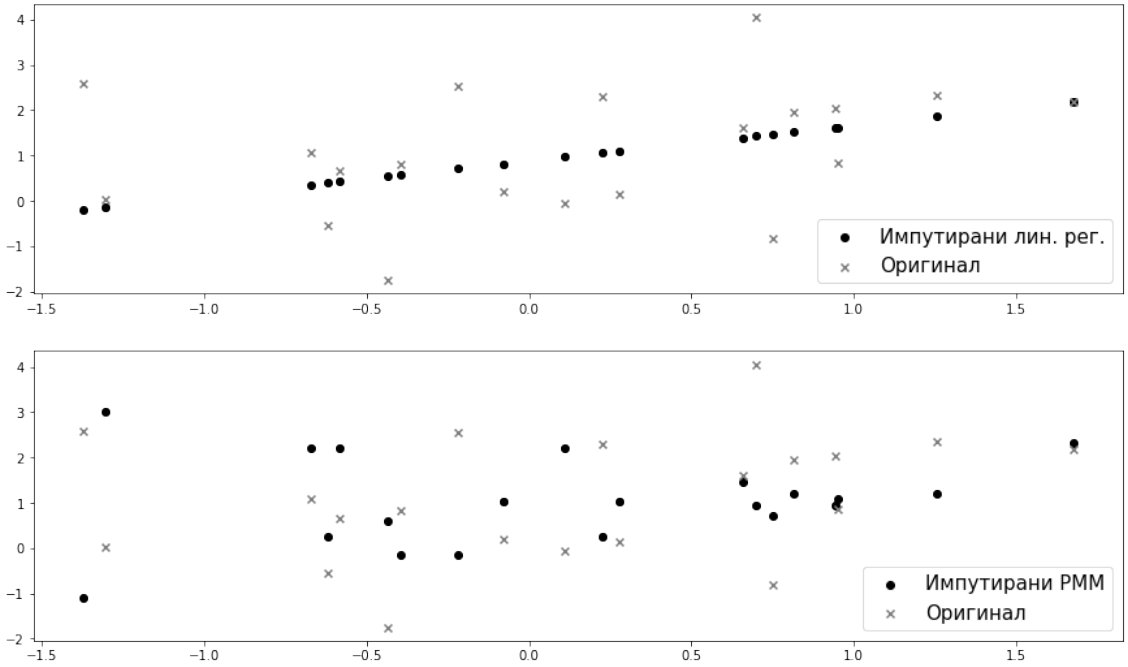
Са слика 3.2 и 3.3 јасно се уочава колико је PMM успешнији у очувању структуре података неголи линеарна регресија. Штавише, већина рачунарских алата за рад са недостајућим подацима при сваком позиву функције која импутира линеарном регресијом кориснику даје упозорење да тај метод не треба користити самостално (то, нпр, ради функција `mice` при позиву помоћи `?mice`).

Пример 3.7 (k најближих суседа)**.** Алгоритам k најближих суседа (енг. *k Nearest Neighbo(u)rs - kNN*) представља један веома сродан алгоритам хот дек фамилији. Нека у подацима Y имамо инстанцу y_i са недостајућим пољем y_{ij}. Налазимо k инстанци најближих инстанци y_j (у односу на неку метрику, која трпи и недостајућа поља) које имају присутно j-то поље и на место непознате вредности y_{ij} импутирамо просек

$$\frac{1}{k} \sum_{t=1}^{k} y_{l_t j},$$

где су $y_{l_1}, \ldots, y_{l_k}$ поменутих k најближих инстанци.

Напомена. У наставку ћемо видети да методи вишеструке импутације често настају модификацијама метода једноструке импутације, па ћемо још примера и једне и друге дати тамо.



Слика 3.3: Поређење линеарно-регресионе и PMM импутације за узорак из дводимензионе нормалне расподеле: импутиране вредности и њихови оригинали

3.5 Закључак

До сада смо теоријски и практично покрили концепт једноструке импутације и размотрили је на неколицини примера. Сада ћемо сумирати закључке које смо у том процесу успут изводили. Литл и Рубин ([5]) од импутационих алгоритама захтевају да буду:

- Условне у односу на доступне податке, да би се смањила пристрасност оцена, побољшала њихова прецизност и очувала веза између доступних и недоступних података која латентно постоји.
- Зависне од више атрибута, да се очува веза међу променљивим у подацима.
- Случајан одабир из неке предиктивне расподеле, пре неголи њено (условно) очекивање, како се не би пренагласила лажна веза међу променљивим.

Основни проблем који имају сви методи једноструке импутације јесте тај што ни на који начин не урачунавају неодређеност[15] коју имамо кад импутирамо. То води ка потцењивању свих стандардних грешака оцена, p-вредности тестова постају превише значајне, а интервали поверења сувише уски. Једини тренутак код којег смо имали покушај да се то санира био је једнодимензиони хот дек, а то је јако примитиван метод. Просто, ми импутирамо недостајуће вредности и онда са таквом базом података настављамо анализу као да је она комплетна. Ма колико квалитетан и напредан био наш импутациони алгоритам, то је очигледно погрешно, јер не урачунава поменуту неодређеност коју смо о недостајућим вредностима имали пре импутације. Можемо рећи да пре импутације о тим вредностима нисмо знали ништа осим да недостају, док након импутације знамо нешто, а понашамо се као да знамо све. Ту је грешка, а ту грешку се труди, и у великој мери успева, да исправи вишеструка импутација, којом се бавимо у наредној глави.

[15] Енг. *uncertainty*. Може и несигурност, неизвесност. Не треба мешати са математичким појмом неодређености (ентропије), иако су очигледно у ускоj вези. Ми ћемо тај термин користити доста слободно.

Глава 4

Вишеструка импутација

4.1 Разни погледи на урачунавање неодређености

У досадашњем раду имали смо ситуације у којима смо на основу импутираних података рачунали оцене параметара, а затим на основу тих оцена изводили извесне закључке. Видели смо на многим примерима како су ти закључци у већини случајева погрешни, јер ниједан метод импутације није у стању да урачуна неодређеност коју имамо о недостајућој вредности.

Деценијском применом разних метода искристалисала су се четири генерална приступа за урачунавање неодређености услед недостајућих података:

1. Да се експлицитно изведу формуле за дисперзију оцене у којима ће се јасно уочити почетна дисперзија коју је оцена имала, а посебно „додатни део". Тада тај додатни сабирак заправо представља пораст дисперзије оцене за који је заслужно присуство недостајућих података. Рецимо, такво једно извођење имали смо у примеру 3.4 где смо изводили дисперзију хот дек оцене за једнодимензионе податке, где је сам узорак ПСУ са понављањем, а извлачењу импутационих вредности такође смо придружили ПСУ без понављања. У том случају били смо у стању да изведемо конкретну дисперзију и видели смо њен пораст услед присуства недостајања.

   Међутим, оваква експлицитна извођења постају компликована чак и за једноставне методе попут хот дек импутације за једнодимензиони узорак, и генерално правци истраживања не иду на ту страну. Ми се овим приступом више нећемо бавити, а у претходној глави имао је само илустративну намену, у циљу разумевања проблема.

2. Да се сам алгоритам којим се импутира модификује тако да се валидна оцена дисперзије може извући из само једне попуњене базе. Овакви приступи су могући и срећу се по литератури, али их је јако тешко генерализати и везани су само за конкретне ситуације и потребе. Такође, жеља за валидном оценом дисперзије често у тој мери „поквари" оцену тако да њена пристрасност неприхватљиво скочи. Нешто више о овоме може се прочитати у [5], а ми му нећемо посвећивати пажњу јер имамо жељу за много општијом теоријом.

3. Да се врши реузорковање доступних података, и онда да се на основу сваког од тих „подузорака" формира оцена жељеног параметра. Варијабилност међу тим оценама тада је одраз неодређености недостајућих вредности. Најчешће се користе бутстреп и џекнајф методи реузорковања. Основна мана овог приступа је та што захтева велики обим узорка за задовољавајуће резултате и веома је рачунски захтеван. И о овоме се може прочитати у [5].

4. *Вишеструка импутација.* Креира се више верзија импутираног скупа података, тако што се различите вредности импутирају на упражњена места. Затим се комбинују оцене дисперзије на основу сваког од тих скупова, са варијабилношћу оцене параметра на сваком од њих и тако се добија валидна оцена дисперзије. Вишеструка импутација је рачунски много мање захтевна и од џекнајфа и од бутстрепа, те се чешће и користи у озбиљним анализама. О њој ћемо у наставку рећи највише.

4.2 Вишеструка импутација: историјски преглед

Пре неголи се значајно формалније посветимо вишеструкој импутацији, рећи ћемо нешто и о њеном настанку. Немамо амбицију да дамо свеобухватан преглед литературе, а заинтересовани га могу пронаћи у [6].

Године 1977. Фриц Шурен радио је у Бироу за попис Сједињених Америчких Држава као аналитичар података. Биро је тада користио (а користи и данас) изворни облик хот дек импутације за попуњавање недостајућих података. Шурен је уочио да при таквој импутацији дисперзије оцена параметара не могу бити правилно срачунате (у огромном броју случајева) и замолио је Доналда Рубина за помоћ. Рубин је дошао на идеју о коришћењу више импутираних база података, те оцењивањем параметара и рачунањем дисперзије на свакоме од њих. Разлике међу тим оценама одражавале би неодређеност коју имамо о недостајућим вредностима. У оригиналном Рубиновом извештају тада није било правила за обједињавање тих оцена, која ће тек касније бити уведена и названа Рубиновим правилима. О томе ћемо у наставку рећи више.

Рубин је ову идеју начелно разматрао у раним седамдесетим, а била је **укорењена у бајесовском начину мишљења**. Скуп различитих оцена параметра (или параметара) може се замислити као извесна апостериорна расподела, из које се касније за оцену параметра узима извесна средња вредност. Математичка позадина вишеструке импутације уследила је 1987. године, такође од стране Рубина, а дата је у књизи [28].

Данас је вишеструка импутација општеприхваћена у научној заједници, и углавном служи као референтни метод са којим се пореде новонастали методи.[1]

4.3 Спецификација бајесовског модела за параметар

Нека нам је Q параметар о којем желимо да изводимо закључке. За сада ћемо претпоставити да је он скаларан, то јест димензије један. Он може бити параметар модела из којег су узорковани подаци, нека функција параметра или параметара модела, или нешто сасвим треће.[2] Уопштено говорећи, о параметрима модела има се неко предзнање, тако да можемо претпоставити да имамо неку информацију о априорној расподели тог параметра, па претпостављамо да је спецификован бајесовски модел

$$f_{[\mathbf{Y},Q]}(\mathbf{y},q) = f_{[\mathbf{Y}|Q]}(\mathbf{y} \mid q) f_Q(q) = f_{[Q|\mathbf{Y}]}(q \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}), \tag{4.1}$$

где нам је $f_{\mathbf{Y}}(\mathbf{y})$ ништа друго до $f_{\boldsymbol{\theta}}(\mathbf{y})$, али смо овде искористили другу ознаку због конзистентности записа. Нагласимо да индексе често нећемо писати због прегледности, осим кад постоји опасност од забуне.

4.3.1 Последица RMAR недостајања

За почетак ћемо формулисати и доказати једну теорему, која на тренутак може деловати невезано за причу коју смо претходно причали, али ће врло брзо добити свој смисао.

Теорема 4.1. ***Претпоставимо да је недостајање RMAR Тада је:***

$$f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) = f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})). \tag{4.2}$$

Другим речима, расподела недостајућих података независна је од механизма недостајања, при услову доступних података, ако важи RMAR услов.

Напомена (конвенција). У наставку текста појављиваће нам се подаци чији је реализовани део једнак $o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$, док је недостајући део једнак неком $\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})$. Згодно би било да имамо ознаку за податке који настају „спајањем" ова два. Договоримо се да такве податке означавамо са $[o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})]$. Пређимо на доказ теореме.

[1] Бурен, [6].

[2] Рецимо, $\mathbf{P}\{\|\mathbf{Y}\| > 0\}$ је параметар, а не мора нужно бити параметар расподеле података.

Доказ. Одмах можемо рачунати да је:

$$
\begin{aligned}
&f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) \\
&\quad= \frac{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}), o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}{f(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})} \\
&\quad= \frac{g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid [o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})]) f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])}{\int_{\{\mathbf{y}\mid o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}} g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid [o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})]) f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])\mathrm{d}\mathbf{y}} \\
&\quad= \frac{g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})) f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])}{g_{\boldsymbol{\phi}}(\widetilde{\mathbf{r}} \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})) \int_{\{\mathbf{y}\mid o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}} f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])\mathrm{d}\mathbf{y}} \\
&\quad= \frac{f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])}{\int_{\{\mathbf{y}\mid o(\mathbf{y},\widetilde{\mathbf{r}})=o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\}} f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])\mathrm{d}\mathbf{y}} \\
&\quad= \frac{f_{\boldsymbol{\theta}}([o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})])}{f(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}))} \\
&\quad= \frac{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})) f(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}))}{f(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}))} \\
&\quad= f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})).
\end{aligned}
$$

Трећа једнакост важи јер је недостајање RMAR па из услова у g могу да се склоне недостајући подаци. Остале једнакости су очигледне, те смо завршили доказ. ■

Напомена (суштина). Шта смо ми доказали? Ми смо, заправо, доказали да вероватноћа са леве стране у изразу (4.2) зависи од $\widetilde{\mathbf{r}}$ **само кроз доступне податке**, не и директно; то су, сложићемо се, различите ствари. Да бисмо боље сагледали последице овог резултата, размотримо следеће. Имамо реализоване податке $\widetilde{\mathbf{y}}$ и реализовани образац недостања $\widetilde{\mathbf{r}}$. Присетимо се, доступне и недоступне податке посматрали смо као векторе, а то смо могли захваљујући изоморфности простора вектора и матрица. Резултат (4.2) заправо каже да ако узмемо неки други образац недостајања, али такав да има **исти број јединица** као $\widetilde{\mathbf{r}}$, расподела недостајућих података условљена њиме и доступним подацима биће иста. Другим речима, под RMAR претпоставком, **докле год имамо исти удео недостајућих података, ма како распоређених, њихова расподела неће зависити од обрасца недостајања**. Ово нам омогућава да приликом моделовања расподеле за $\overline{o}(\mathbf{Y},\mathbf{R})$ у услов можемо да ставимо само доступне податке, игноришући потпуно како је до недостајања дошло.

Напомена (други угао гледања). Теорема 4.1 важи и при бајесовској природи параметара $\boldsymbol{\theta}$ и $\boldsymbol{\phi}$, то јест кад им се придружи априорна расподела. Прошао би исти доказ, само би се у сваком бројиоцу и имениоцу проинтегралило по оба параметра. Неопходно би било додати и претпоставку о априорној независности $\boldsymbol{\theta}$ и $\boldsymbol{\phi}$. Ипак, нама тај контекст неће бити интересантан.

4.4 Бајесовско оправдање вишеструке импутације

Нешто раније спецификовали смо бајесовски модел са којим радимо, а нарочит фокус стављамо на апостериорну расподелу $f(q \mid \mathbf{y})$ параметра од интереса при услову података. При стандардним условима, где не постоје недостајући подаци, није неразумно за оцену параметра Q и оцену дисперзије те оцене узети апопстериорно очекивање $\mathbf{E}(Q \mid \widetilde{\mathbf{y}})$ и дисперзију $\mathbf{D}(Q \mid \widetilde{\mathbf{y}})$ параметра при услову реализованих података. Међутим, ми приступ целим подацима немамо: све што ми о реализованим подацима знамо можемо да закључујемо из пара $\big(o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})\big)$. Стога, за оцену параметра и њену дисперзију ми ћемо (**морамо**) узети

$$\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) \quad \text{и} \quad \mathbf{D}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}).$$

Напомена. Може бити да о параметру стварно нешто знамо па му можемо придружити неку смислену апропрну расподелу. Ипак, углавном о њему нећемо знати много, па је цела бајесовска процедура само формална како бисмо строго математички оправдали наше закључке. То ће нам нешто касније омогућити да бирамо априорну расподелу у складу са потребама рачуна.

Ми наравно расподелу $[Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}]$ не знамо, па не знамо ни њено очекивање и дисперзију. Ипак, можемо срачунати да је

$$f(q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}) = \int_{\{\mathbf{y} \mid o(\mathbf{y}, \widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}})\}} f(q \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}, \widetilde{\mathbf{r}})], \widetilde{\mathbf{r}}) \cdot f(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}) \, \mathrm{d}\mathbf{y}. \tag{4.3}$$

Једначина (4.3) може се користити за симулирање расподеле $f(q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$. Дајмо кратак пример.

Пример 4.1. Уколико је $\mathcal{C}$ неки скуп реалних бројева, онда је, на основу закона великих бројева,

$$\mathbf{P}\{Q \in \mathcal{C} \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}\} = \lim_{m \to +\infty} \frac{1}{m} \sum_{l=1}^{m} \mathbf{P}\{Q \in \mathcal{C} \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}_l, \widetilde{\mathbf{r}})], \widetilde{\mathbf{r}}\},$$

где је $\overline{o}(\mathbf{y}_l, \widetilde{\mathbf{r}})$ l-ти по реду узорак из расподеле $f(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$. Уколико се у горњем изразу „зауставимо" на неком коначном m, такав просек биће квалитетна оцена вероватноће са леве стране.

4.4.1 Како поједноставити?

У општем случају није тачно да је $f(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}) = f(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}))$. Међутим теорема 4.1 даје нам довољне услове под којима то јесте тако, што нам је нарочито битно јер смо онда у стању да градимо модел за узорковање вишеструких импутација само на основу доступних података, игноришући где се они налазе и како су распоређени. Слично, **не мора** бити тачно да је

$$f(q \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}, \widetilde{\mathbf{r}})], \widetilde{\mathbf{r}}) = f(q \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}, \widetilde{\mathbf{r}})]), \tag{4.4}$$

јер на десној страни нема експлицитног условљавања са $\widetilde{\mathbf{r}}$, а разлику та два видели смо у дискусији након теореме 4.1. Модел са леве стране горњег израза ниједан истраживач неће претпоставити; углавном се параметар условљава целим подацима, игноришући који је део настао допуњавањем, а који је оригинално реализован. Ипак, параметар Q је параметар модела **података**, а не недостајања условљеног подацима, па уз стандардну претпоставку **различитости** параметара[3] и овде можемо склонити $\widetilde{\mathbf{r}}$ из услова.

Дефиниција 4.1. Расподелу $f(q \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}, \widetilde{\mathbf{r}})], \widetilde{\mathbf{r}})$ зваћемо *апостериорном расподелом параметра q на основу познато комплетираних података*, а расподелу $f(q \mid [o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \overline{o}(\mathbf{y}, \widetilde{\mathbf{r}})])$ *апостериорном расподелом параметра q на основу непознато комплетираних података.*

Напомена. Леву расподелу у (4.4) Рубин је звао *completed-data distribution*, а десну *complete-data distribution*. Разлика је у томе што је једна у стању да препозна који подаци из услова су аутентични, а који реконструисани, док их друга све посматра једнако.

Напомена. **Ми ћемо $\widetilde{\mathbf{r}}$ и даље наставити да пишемо у услову** да бисмо што општије покрили причу, иако, можда, о неким моделима нећемо знати скоро ништа. Тада треба замислити да уместо њих стоје њихови једноставнији аналогони из овог потпоглавља.

Коначан циљ био нам је да рачунамо (или макар апроксимирамо) $\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$ и $\mathbf{D}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$, па ћемо се сада вратити том проблему. Наравно, дискусија из пододељка 4.4.1 биће нам корисна у даљем раду.

4.4.2 Апостериорно очекивање и дисперзија

Уведимо ознаке за апостерорно очекивање и дисперзију параметра Q на основу познато комплетираних података:

[3] У смислу теореме 2.1 или њихове априорне независности (ако и њима претпоставимо расподеле, зависи од контекста рада) у смислу теореме 2.2.

$$\mathbf{E}(Q \mid [o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})],\widetilde{\mathbf{r}}) =: \check{Q} \quad \text{и} \quad \mathbf{D}(Q \mid [o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})],\widetilde{\mathbf{r}}) =: \check{U}.$$

Суштински, оба су функције од $o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$, $\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})$ и $\widetilde{\mathbf{r}}$, али то углавном нећемо писати јер би запис постао прекомпликован. Оно што је битно је то да су од три претходно набројане ствари - две фиксне: доступни подаци $o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$ и реализовани образац недостајања $\widetilde{\mathbf{r}}$. Оно што може да се мења јесу недоступни подаци $\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})$, па, суштински, $\check{Q}$ и $\check{U}$ јесу функције недостајућих података. Сада на основу познате формуле[4] за условно очекивање можемо срачунати да је

$$\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) = \mathbf{E}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{Q}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})), \tag{4.5}$$

као и[5]

$$\mathbf{D}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) = \mathbf{E}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{U}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})) + \mathbf{D}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{Q}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})). \tag{4.6}$$

Једначине (4.5) и (4.6) веома су корисне јер дају упутство како се могу симулирати апостериорно очекивање и дисперзија који су њима дати.

4.4.3 Симулација апостериорног очекивања и дисперзије

Претпоставимо да смо из расподеле $f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})$ узорковали m независних „допуњавања" података, која ћемо означити са $\overline{o}(\mathbf{y}_l,\widetilde{\mathbf{r}})$ и да смо на основу њих срачунали вредности $\check{Q}$ и $\check{U}$, узевши као фиксне $o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$ и $\widetilde{\mathbf{r}}$. За свако $l = 1, 2, \ldots, m$ означимо

$$\check{Q}_l = \check{Q}(\overline{o}(\mathbf{y}_l,\widetilde{\mathbf{r}})) \quad \text{и} \quad \check{U}_l = \check{U}(\overline{o}(\mathbf{y}_l,\widetilde{\mathbf{r}})).$$

Уколико пустимо да m тежи бесконачности, на основу закона великих бројева[6] у лимесу ћемо добити да је

$$\overline{Q}_\infty := \lim_{m\to\infty}\frac{1}{m}\sum_{l=1}^{m}\check{Q}_l = \mathbf{E}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{Q}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})), \tag{4.7}$$

затим

$$\overline{U}_\infty := \lim_{m\to\infty}\frac{1}{m}\sum_{l=1}^{m}\check{U}_l = \mathbf{E}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{U}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})), \tag{4.8}$$

те коначно

$$B_\infty := \lim_{m\to\infty}\frac{1}{m}\sum_{l=1}^{m}(\check{Q}_l - \overline{Q}_\infty)^2 = \mathbf{D}_{f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}})\mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})}\check{Q}(\overline{o}(\mathbf{Y},\widetilde{\mathbf{r}})). \tag{4.9}$$

Када ово спојимо са (4.5) и (4.6), закључујемо да је

$$\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) = \overline{Q}_\infty \tag{4.10}$$

и

$$\mathbf{D}(Q \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}}) = T_\infty, \tag{4.11}$$

где је $T_\infty = \overline{U}_\infty + B_\infty$.

[4] $\mathbf{E}X = \mathbf{E}(\mathbf{E}(X \mid Y))$.
[5] На основу формуле $\mathbf{D}X = \mathbf{E}\mathbf{D}(X \mid Y) + \mathbf{D}\mathbf{E}(X \mid Y)$.
[6] Осим у (4.9), али је постојаност узорачке дисперзије позната чињеница у статистици.

4.4.4 Симулација за коначно m

До сада смо видели како је могуће изразити $\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$ и $\mathbf{D}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$ у случају када имамо бесконачно много импутација на располагању. Међутим, то у пракси, наравно, никада није случај, те морамо прилагодити процедуру ситуацијама са коначним (потенцијално малим) бројем импутација. Прво ћемо увести неке нове ознаке. Нека нам је

$$\mathcal{S}_m = \left\{\check{Q}_l, \check{U}_l \mid l = 1, 2, \ldots, m\right\}$$

и нека је $\mathcal{S}_\infty$ његов аналогон за бесконачно много импутација. Уочимо ли да апостериорно очекивање и дисперзија дати изразима (4.10) и (4.11) зависе од $[o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}]$ само кроз елементе $\mathcal{S}_\infty$, или, штавише, преко $\overline{Q}_\infty$, $\overline{U}_\infty$ и B_∞, то их можемо посматрати као

$$\mathbf{E}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}) = \overline{Q}_\infty = \mathbf{E}(Q \mid \overline{Q}_\infty, \overline{U}_\infty, B_\infty),$$

односно

$$\mathbf{D}(Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}) = T_\infty = \mathbf{D}(Q \mid \overline{Q}_\infty, \overline{U}_\infty, B_\infty).$$

Напомена (важна претпоставка). Ми ћемо се повести Рубиновим идејама, те ћемо претпоставити да, поред очекивања и дисперзије које смо извели, можемо претпоставити и нормалност:

$$[Q \mid \overline{Q}_\infty, \overline{U}_\infty, B_\infty] \sim \mathcal{N}(\overline{Q}_\infty, \overline{U}_\infty + B_\infty). \tag{4.12}$$

Оваква претпоставка се обично намеће или за класичну фреквенционистичку оцену, или за апостериорну расподелу при целим подацима. Рубин, пак, тврди да у пракси не постоји значајна разлика, јер ће нормалност важити при веома лабавој класи услова.[7]

Оно што ми желимо није ово што тренутно имамо, а то је апостериорна расподела[8] од Q при услову $\mathcal{S}_\infty$, већ жеља да условимо Q са $\mathcal{S}_m$. План је следећи.

План деловања

Сетимо се како смо дошли довде: тражили смо очекивање расподеле $[Q \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}]$, али нисмо знали, па смо је онда условили са недоступним подацима, те упросечили по расподели недоступних података при услову доступних. Слично ћемо и овде.

Како је наш циљ расподела $[Q \mid \mathcal{S}_m]$, а имамо $[Q \mid \overline{Q}_\infty, \overline{U}_\infty, B_\infty]$, то ћемо ми наћи расподелу од $(\overline{Q}_\infty, \overline{U}_\infty, B_\infty)$ условљену са $\mathcal{S}_m$, па упросечити по тој расподели.

Прво ћемо срачунати класичну фреквенционистичку расподелу (*sampling distribution*) елемената из $\mathcal{S}_m$ при услову $o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}$. Уочимо да су $\check{Q}_l$ и $\check{U}_l$ узорковани из расподеле апостериорног очекивања, односно дисперзије (насталих попуњавањем на основу расподеле $f(\overline{o}(\mathbf{y}, \widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}})$). За такве узорке је показано да под веома слабим условима регуларности, у које ми сада нећемо залазити, а могу се наћи у [28], одељак 2.10 и тамошњим референцама, важи да апостериорни просеци теже нормалној расподели, и то код нас

$$[\check{Q}_l \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}] \sim \mathcal{N}(\overline{Q}_\infty, B_\infty). \tag{4.13}$$

Нормалност нам гарантују услови које нисмо излагали; очекивање је очигледно. Што се тиче дисперзије, ту једино може доћи до изненађења, јер бисмо можда очекивали да се појави фактор $\overline{U}_\infty$. Ипак, распишемо ли је аналогно формули (4.6), видећемо да је први сабирак нула, јер је дисперзија унутар очекивања нула. Асимптотски резултати из [28] не гарантују нормалност за $\check{U}_l$, али дају да ће очекивање бити једнако $\overline{U}_\infty$, а дисперзија барем за ред величине мања[9] од дисперзије очекивања. Договоримо се да то означимо као Рубин:

$$[\check{U}_l \mid o(\widetilde{\mathbf{y}}, \widetilde{\mathbf{r}}), \widetilde{\mathbf{r}}] \sim (\overline{U}_\infty, \ll B_\infty). \tag{4.14}$$

[7] Рубин, [28].

[8] Или макар њено очекивање и дисперзија.

[9] У смислу да ће количник да им тежи нули при повећању обима узорка.

Извођење

Прихватимо ли апроксимације (4.13) и (4.14), лако ћемо извести апостериорне расподеле које тражимо. Наиме, ми $\overline{U}_\infty$ можемо посматрати као параметар, те му доделити довољно неинформативну априорну расподелу[10] тако да буде

$$[\overline{U}_\infty \mid \mathcal{S}_m, B_\infty] \sim (\overline{U}_m, \ll B_\infty/m), \tag{4.15}$$

где дељење дисперзије са m долази од упросечавања, а означили смо

$$\overline{U}_m = \frac{1}{m}\sum_{l=1}^{m} \check{U}_l.$$

Слично, ако за $\overline{U}_\infty$ наметнемо (неправу) расподелу пропорционалну константи, имаћемо апостериорну:

$$[\overline{Q}_\infty \mid \mathcal{S}_m, B_\infty] \sim \mathcal{N}(\overline{Q}_m, B_\infty/m), \tag{4.16}$$

где је, јасно

$$\overline{Q}_m = \frac{1}{m}\sum_{l=1}^{m} \check{Q}_l.$$

Сада смо спремни да изведемо апостериорну расподелу параметра Q при услову $\mathcal{S}_m$ и B_∞. Параметар B_∞ нећемо апроксимирати сада на овај начин, већ нешто касније. Једначина (4.15) говори нам да у (4.12) $\overline{U}_\infty$ може бити замењено са $\overline{U}_m$, па ћемо добити (апроксимацију)

$$[Q \mid \mathcal{S}_m, \overline{Q}_\infty, B_\infty] \sim \mathcal{N}(\overline{Q}_\infty, \overline{U}_m + B_\infty). \tag{4.17}$$

Применимо ли сада на ово (4.16), добићемо да је

$$[Q \mid \mathcal{S}_m, B_\infty] \sim \mathcal{N}(\overline{Q}_m, \overline{U}_m + B_\infty + B_\infty/m). \tag{4.18}$$

Напомена (објашњење). Ако је $X \sim \mathcal{N}(\overline{Q}_\infty, \overline{U}_m + B_\infty)$, онда се може записати да је

$$X = \overline{Q}_\infty + \mathcal{N}(0, \overline{U}_m + B_\infty),$$

где је други сабирак формална ознака за случајну величину. Међутим, ми можемо заменити и $\overline{Q}_\infty$ случајним елементом из расподеле $\mathcal{N}(\overline{Q}_m, B_\infty/m)$, па ће се дисперзије сабрати (познато својство нормалне расподеле), те нам дати (4.18).

У једначини (4.18) још је $\overline{B}_\infty$ остало незамењено својом оценом. Јасно, њега ћемо заменити његовом постојаном оценом која ће, наравно, бити узорачка дисперзија:

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m} (\check{Q}_l - \overline{Q}_m)^2,$$

што ће комплетирати наше апроксимације. Да ли ће и како оваква неоправдана замена „покварити" расподелу - видећемо нешто касније.

Рубинова правила

У овом под-пододељку сумираћемо оно што смо извели до сада. Дакле, за оцену параметра Q, у смислу оцене његове апостериорне дисперзије при услову доступних података (и евентуално реализованог механизма недостајања) предлаже се

$$\overline{Q}_m = \frac{1}{m}\sum_{l=1}^{m} \check{Q}_l, \tag{4.19}$$

[10] Енг. *diffuse prior*. О оваквим расподелама може се прочитати у [18] и [16], а ми смо о њима доста научили из [3]. Једна позната је Џефрисова.

а за оцену дисперзије, у смислу апостериорне дисперзије параметра при истом услову, предлаже се

$$T_m = \overline{U}_m + \left(1 + \frac{1}{m}\right) B_m. \tag{4.20}$$

Једначине (4.19) и (4.20), као и разне њихове модификације, познате су под називом *Рубинова правила.*

Напомена (опрез). У претходној дискусији многи закључци и олако замењивање параметара њиховим оценама почивали су на претпоставци нормалности. Она је честа[11] али не и свеприсутна. Због тога се саветује нарочит опрез, то јест провера претпоставке нормалности кад год да је то могуће.

Напомена (занимљивост). Први резултати (пре 1986) нису садржали корекциони фактор $1 + 1/m$ испред B_m у изразу за оцену дисперзије (4.20). Није требало да га садржи ни у Рубиновом и Шенкеровом раду [29], али је анонимни рецензент уочио грешку која је исправљена пре објављивања. Општој научној јавности није познато његово име.

4.4.5 Извори варијабилности

Након низа нажалост неопходних претпоставки[12] оцену апостериорне дисперзије рачунали смо по формули (4.20). Математички, користећи се бајесовским оквиром закључивања, оправдали смо процедуру.

Ипак, због очигледног нагиба ове гране статистике ка практичној примени, није згорег дати и интуитивну представу сваког од сабирака. Томе ћемо посветити ово потпоглавље. Генерално, апостериорна дисперзија долази из три извора:

- $\overline{U}_m$ представља оцену удела дисперзије који је последица тога што смо узели узорак;[13]
- B_m је оцена удела дисперзије који је последица тога што узорак на основу кога оцењујемо није комплетан, већ је и један његов део оцењен;
- B_m/m је корекциони фактор који је последица тога што је $\overline{Q}_\infty$ замењено оценом на основу коначног m, што је повећало укупан варијабилитет.

Последњи сабирак је битан за мале вредности m, јер спречава да се дисперзија потцени, док је за велике m занемарљив. У доба настанка теорије овај услов је био кључан, јер рачунари нису били довољно моћни, па је традиционалан савет био да се узме $m = 3$, $m = 5$ или $m = 10$. Са наглим порастом способности модерних рачунара, променила се и пракса. Тренутно се саветује да се m постави на барем 50.[14].

4.4.6 Конгенијалност

Наиме, ми смо се код бајесовског приступа користили тиме да вредност параметра оцењујемо апостериорним очекивањем при доступним подацима, а дисперзију те „оцене" апостериорном дисперзијом. То смо радили тако што смо рачунали вредност оцена $\check{Q}$ и $\check{U}$ у различитим комплетираним подацима. Међутим, ове оцене су по конструкцији нека апостериорна очекивања и дисперзије (при услову комплетниранх података). Како нисмо знали да рачунамо спољно очекивање у (4.5), то нам нико не гарантује да ћемо знати унутрашње. Углавном и нећемо. Штавише, углавном нећемо имати ни приближну претпоставку о томе шта узети за априорну расподелу. Цео процес је био формалан, да бисмо логички дошли до оцена: о бајесовској природи проблема онај који врши анализу можда и не размишља.

Напомена. О честој (скоро увек) дуалности и на известан начин еквивалентности бајесовског и фреквенционистичког приступа може се читати у [28], одељак 2.10.

[11] Рубин, [28].

[12] За сада. Класа претпоставки за коју је процедура адекватна свакако је шира од оне за коју је доказано адекватна. То ипак, и даље, чека неког новог или старог Рубина.

[13] А не знамо популацију у design-based приступу, или не знамо параметар у приступу модела.

[14] Бурен, [6].

Сходно реченом, често се прибегава апроксимацији самих $\check{Q}$ и $\check{U}$, и то неким статистикама $\hat{Q}$ и $\hat{U}$, од којих је прва оцена параметра Q који је сада фиксан, а друга оцена њене дисперзије.

Такође, веома често (скоро увек) приликом допуњавања података не знамо расподелу $f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{r}})$. Претпоставимо услове који нам омогућавају да склонимо $\widetilde{\mathbf{r}}$ из услова. Тада нам је неопходно да знамо „лакшу" расподелу $f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}))$. Међутим, ни за њу нам нико не гарантује да је знамо (углавном је и не знамо). Стога се и она апроксимира неком расподелом $f_I(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{a}})$, где $\widetilde{\mathbf{a}}$ означава реализацију неког додатног случајног елемента о којем онај који импутира нешто зна (на пример неке случајне величине која није међу подацима).

На ово је указао Менг[15], који је увео концепт *конгенијалности*.[16]

Дефиниција 4.2. Бајесовски модел (4.1) је *конгенијалан* са импутационом расподелом и паром[17] $(\hat{Q},\hat{U})$ *при датим доступним подацима* $o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$ уколико важе следећи услови:

1. За било које комплетне податке $\mathbf{y}$ за које је $o(\mathbf{y},\widetilde{\mathbf{r}}) = o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})$ важе једнакости

$$\mathbf{E}_{f(Q|\mathbf{Y})}(Q \mid \mathbf{y}) = \hat{Q}(\mathbf{y}) \quad \text{и} \quad \mathbf{D}_{f(Q|\mathbf{Y})}(Q \mid \mathbf{y}) = \hat{U}(\mathbf{y}). \tag{4.21}$$

2. Импутациона расподела једнака је правој, то јест

$$f_I(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}}),\widetilde{\mathbf{a}}) = f(\overline{o}(\mathbf{y},\widetilde{\mathbf{r}}) \mid o(\widetilde{\mathbf{y}},\widetilde{\mathbf{r}})).$$

Конгенијалност нас заправо ослобађа проблема замене апостериорног очекивања и дисперзије одговарајућим реализацијама статистика $\hat{Q}$ и $\hat{U}$, јер смо при претпоставци конгенијалности сигурни да су они узорковани баш из оне расподеле на чији се закон великих бројева позивамо.

Напомена. Наравно, у пракси је довољно да конгенијалност важи приближно.

Напомена. У последњих неколико година појавило се мноштво нових резултата за генерисање вишеструких импутација у одсуству конгенијалности, а нарочито треба истаћи резултате до којих су дошли Сје (Xie) и Менг, а који се могу наћи у [8]. Ми се њима овде нећемо бавити.

4.5 Фреквенционистичка својства оцена

Извођење статистичких закључака на основу оцена које смо имали у претходним одељцима ове главе почивало је на великом броју претпоставки, као што су разни видови нормалности, разни имплицитни услови регуларности, као и претпоставка да особа која импутира и особа која врши анализу података користе исти бајесовски модел. Ово последње смо елегантно „спаковали" у претпоставку конгенијалности.

Ипак, како често барем нека од ових претпоставки може бити и бива нарушена, пожељно је испитати својства бајесовски добијених оцена на класичан, фреквенционистички начин. Томе ћемо посветити овај одељак.

Напомена. Како ћемо податке и обрасце недостајања овде сматрати случајним, нећемо користити тилду за реализоване податке/обрасце, јер нам је та ознака била уско везана за **дате** податке, оне стварне које имамо у истраживању, а не оне који су „вештачки" узорковани у сврхе евалуације.

4.5.1 Валидност закључивања на основу целих података

Нека и овде имамо параметар Q (и даље скаларан), за који је предложена оцена $\hat{Q}$ и оцена њене дисперзије $\hat{U}$. У претходном поглављу сусрели смо се са појмом конгенијалности, који

[15] Менг, [25]. Ипак, дефиницију у овој форми преузели смо из [8], јер је рад новијег датума те је формулација дефиниције много природнија.

[16] Енг. *congeniality*.

[17] Њега нема у оригиналној дефиницији, али је јасно да га Менг подразумева.

нам је омогућавао да значајно апстрактне појмове попут апостериорног очекивања и дисперзије параметра, које често не знамо да рачунамо, или нам је то рачунски неисплативо, можемо заменити вредношћу статистика $\hat{Q}$ и $\hat{U}$ у комплетираним подацима. Ипак, уколико су саме оцене $\hat{Q}$ и $\hat{U}$, посматране као обичне статистике на комплетним подацима, неквалитетне,[18] нереално је очекивати да вишеструка импутација, или било која техника руковања недостајућим подацима, „излечи" тај проблем. У пракси, наравно, скоро никада и нећемо рачунати апостериорна очекивања и дисперзије параметара, већ ћемо их мењати оценама $\hat{Q}$ и $\hat{U}$ евалуираним баш у тим комплетираним подацима. Због тога, желели бисмо да саме оцене $\hat{Q}$ и $\hat{U}$, у одсуству недостајућих података, буду квалитетне. Да бисмо могли рећи да ли је нека оцена квалитетна или не, морамо задати неке пожељне особине по којима ћемо их класификовати. Ми ћемо сада навести једну од најбитнијих, коју је дефинисао Рубин, у [28].

Дефиниција 4.3. Кажемо да је закључивање на основу комплетних података и на основу статистика $(\hat{Q}, \hat{U})$ *валидно у смислу интервала поверења*[19] уколико важе својства

$$\hat{Q} \sim \mathcal{N}(Q, \mathbf{D}\hat{Q}) \tag{4.22}$$

и

$$\mathbf{E}\hat{U} \geqslant \mathbf{D}\hat{Q}. \tag{4.23}$$

Уколико важи (4.22), уколико у (4.23) стоји једнакост и уколико је

$$\mathbf{D}\hat{U} \leqslant \mathbf{D}\hat{Q}, \tag{4.24}$$

онда кажемо да је закључивање *рандомизационо валидно.*[20]

Напомена (појашњење). Извођење квалитетних интервала поверења један је од основних задатака статистике. Штавише, интервали поверења су неретко уско повезани са критичним областима тестова (комплемент су један другог), што је још један прилог тврдњи да се извођењима интервала поверења мора приступити са нарочитом пажњом. Чест облик интервала поверења јесте

$$\big(\text{ОП} - z \cdot \text{ОДО}, \text{ОП} + z \cdot \text{ОДО}\big),$$

где под ОП подразумевамо оцену неког параметра, а под ОДО оцену дисперзије те оцене. Сада валидност у смислу интервала поверења постаје много јаснија: услов непристрасности, (4.22), говори нам о томе да ће интервал бити добро „центриран", то јест да неће бити систематске пристрасности у његовом позиционирању; услов (4.23) нам, пак, говори да ће интервал бити широк барем онолико колико треба. Другим речима, реална стопа покривености праве вредности параметра може бити само мања од номиналне, никако и већа. Овај услов онемогућава интервал поверења да буде „лажно" узак.

Наравно, најбоље би било да стварна стопа покривености буде баш једнака номиналној, али како је то често нереално очекивати, ипак сматрамо бољим прецењити дисперзију, неголи је потценити.

Такође, у пракси је довољно да нормалност важи приближно.

4.5.2 Валидност закључивања у присуству недостајања: случај бесконачног m

Претпоставимо да на располагању имамо бесконачно много вишеструких импутација и да за анализу користимо статистике $\hat{Q}(\mathbf{Y})$ и $\hat{U}(\mathbf{Y})$. Претпоставимо да се и даље држимо претпоставке о нормалности:

$$[Q \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}] \sim \mathcal{N}(\overline{Q}_\infty, T_\infty).$$

[18] Нпр. пристрасне, непостојане итд.

[19] Енг. *confidence valid.*

[20] Што је строжији услов: рандомизационо валидно закључивање је валидно у смислу интервала поверења, али обрнуто не мора бити.

Оно што треба приметити јесте да су $\overline{Q}_\infty$ и T_∞, редом, апостериорно очекивање и дисперзија расподеле $[Q \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}]$, па су сами функције оног што је у услову расподеле, односно функције од $o(\mathbf{y}, \mathbf{r}), \mathbf{r}$.

Оно што бисмо сада желели да испитамо јесте какав ће бити квалитет ових оцена[21] када се ови аргументи, које смо до сада држали фанатично фиксним, замене случајним елементима и почну мењати. Овде ћемо индикатор доступности, $\mathbf{R}$, звати и индикатор одговора[22]. Следећу дефиницију дао је Рубин 1987.

Дефиниција 4.4. У досадашњим ознакама, вишеструка импутација је *рандомизационо валидна под претпостављеним моделом одговора*[23] $g_\phi(\cdot)$, уколико важи да је

$$\overline{Q}_\infty \sim \mathcal{N}(Q, T_0) \tag{4.25}$$

и[24]

$$T_\infty \sim (T_0, \ll T_0), \tag{4.26}$$

где су Q и T_0 праве вредности одговарајућих параметара.

Напомена. Овде је у позадини заједничка расподела за податке $\mathbf{Y}$ и индикатор $\mathbf{R}$.

Није тешко уочити да ће услови (4.25) и (4.26) важити за велике обиме узорке уколико важе услови конгенијалности то јест уколико су вишеструке импутације узорковане из правог модела недостајућих података при доступним и ако се слажу апостериорна очекивања и дисперзије са статистикама $\hat{Q}$ и $\hat{U}$ које их „мењају".

Напомена (практични проблеми). Један од основних проблема на који се може наићи смо већ поменули, а то је одсуство конгенијалности. Оно је честа појава када особа која импутира није и особа која анализира податке. Други проблем који се може јавити јесте да се стварни и импутациони модел уопште не могу поредити, јер овај други уопште не мора бити експлицитно задат, већ може (и у пракси углавном јесте) задат алгоритмом.

Напомена. Убудуће чемо претпостављати да важи конгенијалност и нећемо разликовати $(\hat{Q}, \hat{U})$ од $(\check{Q}, \check{U})$.

Правилност вишеструке импутације по Рубину

Претпоставке под којима смо извели рандомизациону валидност веома су јаке, па бисмо волели да имамо и општије услове под којима она важи. Сада ћемо дефинисати један од њих.

Дефиниција 4.5. У досадашњим ознакама, вишеструка импутација је *правилна* за пар статистика $(\hat{Q}, \hat{U})$ уколико важе наредна три услова:

1. Сматрајући податке $\mathbf{y}$ фиксним, под претпостављеним моделом одговора, важи да је

$$[\overline{Q}_\infty \mid \mathbf{y}] \sim \mathcal{N}(\hat{Q}(\mathbf{y}), V(\mathbf{y})) \tag{4.27}$$

и

$$[B_\infty \mid \mathbf{y}] \sim (V(\mathbf{y}), \ll V(\mathbf{y})), \tag{4.28}$$

где, подсетимо се, $\ll V$ значи да је дисперзија мањег реда од V, које смо дефинисали као

$$V(\mathbf{y}) = \mathbf{D}\overline{Q}_\infty(\mathbf{y}). \tag{4.29}$$

Сва очекивања и дисперзије узимају се по расподели за $\mathbf{R}$ (то је једино што нисмо фиксирали).

[21] За почетак за бесконачно m, а касније и за коначно.

[22] Историјски, због контекста настанка вишеструке импутације за потребе обраде података насталих анкетирањем.

[23] Који не мора бити онај који је заиста стваран. Ми овакву валидност дефинишемо за претпостављени модел. За неке моделе процедура ће бити валидна, за неке неће.

[24] Подсетимо се, ознака значи да је очекивање расподеле T_0, а дисперзија реда величине мањег од T_0.

2. Сматрајући податке $\mathbf{y}$ фиксним, под претпостављеним моделом одговора, важи да је

$$[\overline{U}_\infty \mid \mathbf{y}] \sim (\hat{U}(\mathbf{y}), \ll V(\mathbf{y})). \tag{4.30}$$

3.
$$\mathbf{D}V(\mathbf{Y}) \leqslant \mathbf{E}\hat{U}(\mathbf{Y}), \tag{4.31}$$

где очекивање и дисперзија иду по расподели за $\mathbf{Y}$.

Наредну теорему дао је Рубин у [28].

Теорема 4.2. ***Уколико је закључивање на основу комплетних података рандомизационо валидно и уколико је вишеструка импутација правилна, онда је вишеструка импутација рандомизационо валидна под претпостављеним моделом одговора, у смислу дефиниције 4.4.***

Доказ. Оно што треба да докажемо јесте да из релација (4.22) - (4.24) и (4.27) - (4.31) следе релације (4.25) и (4.26).

Једнакости (4.22) и (4.27) заједно нам дају да је

$$\overline{Q}_\infty \sim \mathcal{N}(Q, \mathbf{D}\hat{Q}(\mathbf{Y}) + \mathbf{E}V(\mathbf{Y})),$$

где су спољна очекивања и дисперзија по расподели за $\mathbf{R}$. Очекивање је очигледно, док дисперзија следи из већ виђене формуле $\mathbf{D}X = \mathbf{ED}(X \mid Y) + \mathbf{DE}(X \mid Y)$.

Следеће што ћемо да уочимо јесте то да из (4.28) и (4.30) следи да је

$$[\overline{U}_\infty + B_\infty \mid \mathbf{y}] \sim (\hat{U}(\mathbf{y}) + V(\mathbf{y}), \ll 2V(\mathbf{y})),$$

што нам укомбиновано са (4.23) и (4.31) даје

$$\overline{U}_\infty + B_\infty \sim (\mathbf{D}\hat{Q}(\mathbf{Y}) + \mathbf{E}V(\mathbf{Y}), \ll (2\mathbf{E}V(\mathbf{Y})) + 2\mathbf{D}\hat{Q}(\mathbf{Y})),$$

што завршава доказ, због значења ознаке $\ll$. ■

4.5.3 Случај коначног m и асимптотика оцена

Као што смо видели у претходном пододељку, онда када је вишеструка импутација правилна по Рубину за бесконачно m и када је закључивање на основу комплетних података рандомизационо валидно, тада је цела процедура рандомизационо валидна. Наравно, у пракси, ма колико много их имали, никада нећемо имати на располагању бесконачно много импутација, тако да се природно мора поставити и питање фреквенционистичких перформанси на основу коначног броја импутација. Стога, нећемо се задовољити утврђивањем расподеле од $\overline{Q}_\infty$, $\overline{U}_\infty$ и B_∞ при случајном узорковању и недостајању, већ ћемо потражити расподелу од $\overline{Q}_m$, $\overline{U}_m$ и B_m.

У сврхе евалуације оцена, поред неких старих претпоставки, наметнућемо и нове. Узевши заједно, претпостављамо следеће:

1. Вишеструка импутација је правилна, у смислу дефиниције 4.5;
2. Закључивање на основу комплетних података је рандомизационо валидно;
3. Обим узорка је довољно велики да апроксимације за расподеле елемената из $\mathcal{S}_m$ важе, то јест да је

$$[\hat{Q}_l \mid \mathbf{y}, \mathbf{r}] \sim \mathcal{N}(\overline{Q}_\infty, B_\infty) \tag{4.32}$$

и

$$[\hat{U}_l \mid \mathbf{y}, \mathbf{r}] \sim (\overline{U}_\infty, \ll B_\infty), \tag{4.33}$$

где су $\hat{Q}_l$ и $\hat{U}_l$ независни за свако $l = 1, 2, \ldots, m$.

Напомена. У било којој од релација (4.32) и (4.33) у услову, строго теоријски гледано треба да стоји $o(\mathbf{y},\mathbf{r}),\mathbf{r}$, али како ове две оцене од реализованих података зависе само кроз поменуте, свеједно је шта пишемо, па смо се одлучили за једноставнији запис.

Расподеле оцена

Претпоставимо и даље да имамо скаларан параметар. Циљ нам је да изведемо расподеле од $\overline{Q}_m$, $\overline{U}_m$ и B_m. Ово извођење имаће три природна корака:

1. Прво, упросечавамо преко вишеструких импутација фиксирајући податке и недостајање,[25] претпостављајући асимптотску валидност оцена из $\mathcal{S}_m$.

2. Друго, упросечавамо преко расподеле од $\mathbf{R}$, фиксирајући узорак, претпостављајући правилност импутације при претпостављеном моделу недостајања.

3. Треће, упросечавамо преко расподеле података, претпостављајући рандомизациону валидност за закључивање у смислу дефиниције 4.3.

Прођимо кроз процедуру.

1. Из (4.32) и (4.33) одмах можемо закључити да је

$$[\overline{Q}_m \mid \mathbf{y},\mathbf{r}] \sim \mathcal{N}(\overline{Q}_\infty, B_\infty/m), \tag{4.34}$$

$$[\overline{U}_m \mid \mathbf{y},\mathbf{r}] \sim (\overline{U}_\infty, \ll B_\infty/m) \tag{4.35}$$

и, као позната особина узорачке дисперзије, што B_m и јесте, при претпоставци нормалности:

$$\left[\frac{(m-1)B_m}{B_\infty} \,\middle|\, \mathbf{y},\mathbf{r}\right] \sim \chi^2_{m-1}. \tag{4.36}$$

Све три горње случајне величине су међусобно независне. Прве две по конструкцији. B_m не зависи од $\overline{U}_m$ јер зависи само од $\hat{Q}_l$-ова, а не зависи од $\overline{Q}_m$, на основу Басуове теореме, јер је другоразредна статистика за параметар положаја.[26]

2. Пређимо на други корак и фиксирајмо узорак. Скупимо ли релације (4.27), (4.28), (4.34) и (4.36), добијамо да је

$$[\overline{Q}_m \mid \mathbf{y}] \sim \mathcal{N}(\hat{Q}(\mathbf{y}), (1+1/m)V(\mathbf{y})), \tag{4.37}$$

где смо, сетимо се, дефинисали да је $V(\mathbf{y}) = \mathbf{D}\overline{Q}_\infty(\mathbf{y})$, где је у позадини расподела од $\mathbf{R}$. Слично као и малопре,

$$\left[\frac{(m-1)B_m}{V(\mathbf{y})} \,\middle|\, \mathbf{y}\right] \sim \chi^2_{m-1}. \tag{4.38}$$

Као и малопре, спојићемо (4.30) и (4.28) и (4.35) и добијемо да је

$$[\overline{U}_m \mid \mathbf{y}] \sim \big(\hat{U}(\mathbf{y}), \ll (1+1/m)V(\mathbf{y})\big). \tag{4.39}$$

Горње три случајне величине[27] су међусобно независне из истих разлога као у претходној тачки.

[25] Дакле, у позадини је расподела $[\overline{o}(\mathbf{y},\mathbf{r}) \mid o(\mathbf{y},\mathbf{r}),\mathbf{r}]$.
[26] Видети [18].
[27] Или, боље је рећи, расподеле.

3. Коначно, пређемо на трећи корак и посматрамо и податке променљиво. Релације (4.22), (4.23), (4.31) и (4.37) дају да је

$$\overline{Q}_m \sim \mathcal{N}\big(Q, \mathbf{D}\hat{Q}(\mathbf{Y}) + (1+1/m)\mathbf{E}V(\mathbf{Y})\big). \tag{4.40}$$

Затим, једначине (4.40) и (4.38) дају нам, као и до сад, да је

$$\frac{(m-1)B_m}{\mathbf{E}V(\mathbf{Y})} \sim \chi^2_{m-1}. \tag{4.41}$$

Коначно, (4.23), (4.31) и (4.39) дају нам да је

$$\overline{U}_m \sim \left(\mathbf{D}\hat{Q}(\mathbf{Y}), \ll \big(\mathbf{D}\hat{Q}(\mathbf{Y}) + (1+1/m)\mathbf{E}V(\mathbf{Y})\big)\right). \tag{4.42}$$

И овде имамо независност, као и малопре, из сличних разлога.

4.6 Случај вишедимензионог параметра

У свим досадашњим разматрањима смо, поред класичних претпоставки игнорабилности, нормалности, валидности и правилности, претпостављали и да је параметар Q скаларан, односно елемент (неког подскупа) скупа реалних бројева. Међутим, то често неће бити случај и параметар од интереса биће вишедимензион. Рецимо, једнодимензиона нормална расподела јединствено је одређена задавањем параметра очекивања $\mu \in \mathbb{R}$ и параметра дисперзије $\sigma^2 > 0$. У овој ситуацији је $Q = (\mu, \sigma^2)$, што је димензије два. Због оваквих примера, желели бисмо да процедуру вишеструке импутације прилагодимо вишедимензионом Q.

Срећом, то није превише тешко. Наиме, ми смо се при конструкцији оцена највише позивали на формулу $\mathbf{E}X = \mathbf{E}(\mathbf{E}(X \mid Y))$, као и на формулу $\mathbf{D}X = \mathbf{E}\mathbf{D}(X \mid Y) + \mathbf{D}\mathbf{E}(X \mid Y)$. Обе ове формуле важе и у вишедимензионом случају где је просто неопходно заменити дисперзију коваријационом матрицом. Слично је и са претпоставком нормалности, коју природно мењамо претпоставком вишедимензионе нормалности.

Оно што ће се разликовати јесте употреба ознаке $M \sim (M_0, \ll N)$, која је до сада означавала да је дисперзија од M мањег реда величине него N. Како је у овом контексту M вектор, а N матрица, ову ознаку тумачимо тако да је дисперзија сваке компоненте у M мањег реда величине од сваке позитивне компоненте у N.

Још једна разлика јесте у расподели B_m, попут оне у једнакости (4.41). У том случају не можемо делити дисперзијом, јер је она сада матрица, већ ћемо просто множити њеним инверзом. Расподела тада престаје да буде хи-квадрат, и постаје тзв. Висхартова[28] расподела са $m-1$ степени слободе. Сви остали резултати стоје.

4.7 Интервалне оцене и тестови: уопштено у присуству недостајућих података

4.7.1 Конструкције при различитим видовима закључивања

Претпоставимо да имамо k-димензионални параметар Q. Под интервалном оценом параметра Q подразумевамо неки подскуп простора $\mathbb{R}^k$ такав да

(а) зависи само од доступних вредности узорка и

(б) садржи Q са извесном стопом покривања $1-\alpha \in [0,1]$.

[28] Добија се слично као и хи-квадрат, само се не сабирају квадрати нормално расподељених случајних величина, већ њихови умношци својим транспонатом.

Под стопом покривања подразумеваћемо различите ствари у зависности од вида статистичког закључивања које користимо. Рецимо, за фреквенционистичко закључивање сам интервал је случајан, и стопа покривања представља вероватноћу да тај случајан интервал обухвати неслучајан параметар. Код бајесовског закључивања, пак, интервал је неслучајан, а параметар случајан, па стопа покривања представља интеграл преко интервала густине расподеле параметра (или суму вредности закона расподеле у дискретном случају).

Ипак, у пракси се јасно искристалисао начин за изградњу интервала поверења. Генерално, процедура се састоји из

(а) Тачкасте оцене параметра: $\hat{Q}(o(\mathbf{y},\mathbf{r}),\mathbf{r})$;

(б) Статистике која оцењује дисперзију од $\hat{Q}-Q$, коју ћемо означити са $T(o(\mathbf{y},\mathbf{r}),\mathbf{r})$;

(в) Претпоставке нормалности:

$$\hat{Q}-Q\sim\mathcal{N}(0,T).$$

Евентуално, у специфичним ситуацијама дâ се претпоставити нека друга расподела.

На овај начин интервали се конструишу и за бајесовски и за фреквенционистички вид закључивања, само се очекивања и дисперзије у једном случају узимају по параметру, а у другом по подацима/недостајању. Дакле, генерално говорећи, ако је $\mathcal{C}=\mathcal{C}(o(\mathbf{y},\mathbf{r}),\mathbf{r})$ интервална оцена параметра Q, онда номинална стопа покривања $1-\alpha$ мора задовољити једнакост

$$\mathbf{P}\{Q\in\mathcal{C}\mid\mathfrak{A}\}=1-\alpha, \tag{4.43}$$

где је $\mathfrak{A}$ скуп услова којим условљавамо: код бајесовског закључивања то је $\{o(\mathbf{y},\mathbf{r}),\mathbf{r}\}$, а код фреквенционистичког то је евентуално вредност реализованог обрасца недостајања, $\mathbf{r}$, или чак ништа.

Напомена. Иако, у начелу, интервалне оцене не морају представљати математички појам интервала из историјских разлога користићемо тај назив.

4.7.2 Фреквенционистичке интервалне оцене

У овом пододељку размотрићемо интервалне оцене код којих је сам интервал случајан, а параметар неслучајан. Уочимо да у тој ситуацији две ствари могу бити случајне: индикатор одговора $\mathbf{R}$ и узорак $\mathbf{Y}$, па ћемо разматрати разне типове интервала у зависности од тога шта од тога сматрамо фиксним (ставимо у услов), а шта случајним.

Фреквенционистички интервали при фиксном одговору

У досадашњим ознакама, стопа покривања интервалне оцене $\mathcal{C}=\mathcal{C}(o(\mathbf{y},\mathbf{r}),\mathbf{r})$ може се рачунати као

$$\mathbf{P}\big\{Q\in\mathcal{C}\mid\mathbf{r}\big\}=\mathbf{E}\big[I\{Q\in\mathcal{C}\}\mid\mathbf{r}\big]=\int I\{Q\in\mathcal{C}\}f(\mathbf{y}\mid\mathbf{r})\,\mathrm{d}\mathbf{y}. \tag{4.44}$$

Горња вероватноћа зове се *рандомизациона покривеност параметра Q интервалом $\mathcal{C}$ при фиксном недостајању* (или фиксном одговору). Оно што је код оваквих интервала случајно јесу подаци $\mathbf{Y}$, па кроз њих и сам интервал $\mathcal{C}$. Параметар је фиксан.

Фреквенционистички интервали при променљивом одговору

Згодно је имати и фреквенционистичку интервалну оцену која у своје моделовање експлицитно укључује и недостајање $\mathbf{R}$, онда када се сматра да ће оно значајно допринети квалитету модела. Тада је стопа покривања дата са

$$\mathbf{P}\big\{Q\in\mathcal{C}\big\}=\mathbf{E}\big[I\{Q\in\mathcal{C}\}\big]$$

$$
\begin{aligned}
&= \iint I\{Q \in \mathcal{C}\} f(\mathbf{y}, \mathbf{r}) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{r} \\
&= \iint I\{Q \in \mathcal{C}\} f(\mathbf{y} \mid \mathbf{r}) f(\mathbf{r}) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{r} \\
&= \int \left[\int I\{Q \in \mathcal{C}\} f(\mathbf{y} \mid \mathbf{r}) \, \mathrm{d}\mathbf{y} \right] f(\mathbf{r}) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{r} \\
&= \int \mathbf{P}\big\{Q \in \mathcal{C} \mid \mathbf{r}\big\} f(\mathbf{r}) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{r}. \qquad (4.45)
\end{aligned}
$$

Горња вероватноћа зове се *рандомизациона покривеност параметра Q интервалом $\mathcal{C}$ при случајном недостајању*. Видимо да је она једнака просеку стопа покривања при фиксном недостајању.

Напомена. Често разликујемо *номиналну* (теоријску) и *стварну* стопу покривања. Разлог томе су успутне апроксимације које често „покваре" првобитне претпоставке, па стопа покривања буде смањена. Рубин је[29] уочио да вишеструка импутација понекад може да измени интервале до те мере да они не постану ништа шири, а да имају стварну стопу покривања већу од номиналне. Он је овакву појаву назвао *суперефикасност*. Ми се њоме нећемо бавити.

4.7.3 Бајесовске интервалне оцене

Постоје два типична начина за конструкцију интервалних оцена при бајесовској природи параметра Q:

- Да се унапред фиксира стопа покривања $1-\alpha$ и да се нађе $\mathcal{C} \subseteq \mathbb{R}^k$ који укључује оне вредности Q где је апостериорна густина највећа. Често је претходно потребно дефинисати форму $\mathcal{C}$ да би се процедура могла спровести.
- Да се фиксира вредност Q_0 параметра Q, коју ћемо звати *нулта вредност*, те да се $\mathcal{C}$ одреди као скуп вредности параметра Q вероватнијих од Q_0, где се под „вероватнијим" подразумевају оне вредности у којима је густина већа неголи у нултој. Онда се стопа покривања рачуна интегралећи густину преко области $\mathcal{C}$.

Прве ћемо описати мало детаљније.

Региони највеће апостериорне густине

Нека смо унапред фиксирали стопу покривања $1-\alpha$. Да би област $\mathcal{C}$ била интервална оцена за параметар Q са стопом покривања $1-\alpha$ захтеваћемо да важе две ствари:

- Апостериорна вероватноћа да се параметар Q налази у $\mathcal{C}$ једнака је $1-\alpha$:

$$\mathbf{P}\big\{Q \in \mathcal{C} \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}\big\} = \int I\{Q \in \mathcal{C}\} f(Q \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}) \, \mathrm{d}Q = 1 - \alpha. \qquad (4.46)$$

- Вредност апостериорне густине у свакој тачки из $\mathcal{C}$ већа је од вредности апостериорне густине у било којој тачки ван $\mathcal{C}$:

$$(\forall Q' \in \mathcal{C})(\forall Q'' \ni \mathcal{C}) \; f(Q' \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}) > f(Q'' \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}).$$

Напомена (нотација). Због скраћивања записа, поменуту апостериорну густину означићемо као

$$d(Q) := f(Q \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}),$$

подразумевајући шта је у услову.

[29] Рубин, [24].

Познат је резултат да уколико је апостериорна расподела за Q k-димензиона (регуларна) нормална $\mathcal{N}(\hat{Q}, T)$ расподела, да онда важи једнакост у расподели:

$$(Q-\hat{Q})T^{-1}(Q-\hat{Q})^T \sim \chi^2_k.$$

Подсетимо се да је Q врста-вектор.

Стога није тешко конструисати област највеће апостериорне густине као скуп свих Q за које је

$$(Q-\hat{Q})T^{-1}(Q-\hat{Q})^T < \chi^2_k(\alpha),$$

где смо увели скраћеницу $\chi^2_k(\alpha) = F^{-1}_{\chi^2_k}(1-\alpha)$.

Општије, уколико претпоставимо да $Q-\hat{Q}$ има апостериорну Студентову t расподелу са ν степени слободе матрицом скалирања $T^{1/2}$ (која постоји због регуларности и симетричности матрице T), онда се регион највеће апостериорне густине налази као скуп свих Q за које важи

$$(Q-\hat{Q})T^{-1}(Q-\hat{Q})^T < kF_{k,\nu}(\alpha),$$

где је $F_{k,\nu}(\alpha) = F^{-1}_{F_{k,\nu}}(1-\alpha)$, а $F_{k,\nu}$ је Фишерова F расподела са k и ν степени слободе. Овај резултат нећемо доказивати.[30]

4.7.4 *p*-вредности

Претпоставимо да желимо да тестирамо нулту хипотезу $H_0 : Q = Q_0$, против неке од могућих алтернатива, са нивоом значајности α. Желели бисмо да на основу доступних података сумирамо „доказе" који се тичу ове тврдње.

Са бајесовске тачке гледишта ово се може урадити тако што се нађе регион $\mathcal{C}$ највеће апостериорне густине, и да се констатује да ли Q_0 припада овом региону. Ако припада, онда ћемо остати при тврдњи H_0, а иначе ћемо је одбацити. Ипак, постоји мало мање „наиван" начин да се то уради, а он не захтева спецификацију нивоа α унапред. Изделићемо скуп свих могућих вредности за Q на два скупа:

- Скуп оних Q у којима је апостериорна густина већа неголи у Q_0;
- Скуп оних Q у којима је апостериорна густина мања или једнака неголи у Q_0.

Затим срачунамо апостериорну вероватноћу да се Q нађе у првој групи и означимо је са $1-\alpha$. Што је ова вероватноћа већа, то је вредност Q_0 „мање вероватна". Стога се α назива и *бајесовски ниво значајности* нулте вредности Q_0, а понекад и p-вредност од Q_0. Ова вредност заправо представља вероватноћу да параметар узоркован из апостериорне расподеле буде „екстремнији" од Q_0, што одговара класичном појму p-вредности. Записано формално:

$$\textbf{p-val}\{Q_0 \mid o(\mathbf{y},\mathbf{r}),\mathbf{r}\} = \mathbf{P}\{d(Q) < d(Q_0) \mid o(\mathbf{y},\mathbf{r}),\mathbf{r}\}.$$

Специјално, при претпоставци апостериорне нормалности, можемо срачунати да је

$$\textbf{p-val}\{Q_0 \mid o(\mathbf{y},\mathbf{r}),\mathbf{r}\} = \mathbf{P}\{\chi^2_k > (Q-\hat{Q})T^{-1}(Q-\hat{Q})^T\}.$$

Слично, при апостериорној студентовости са ν степени слободе је

$$\textbf{p-val}\{Q_0 \mid o(\mathbf{y},\mathbf{r}),\mathbf{r}\} = \mathbf{P}\{kF_{k,\nu} > (Q-\hat{Q})T^{-1}(Q-\hat{Q})^T\}.$$

[30] Погледати [28], страна 60, и тамошње референце.

4.7.5 Евалуација процедура за конструкцију интервалних оцена

Претпоставимо да је $\mathcal{C} = \mathcal{C}(o(\mathbf{y}, \mathbf{r}), \mathbf{r})$ интервална оцена за параметар Q. Вероватноћа да јој Q припада може бити посматрана на три начина:

- Параметар има (апостериорну) расподелу и номинална стопа прекривања је $1 - \alpha_b$.
- Како је $\mathcal{C}$ функција доступних података и реализованог обрасца недостајања, фиксираћемо $\mathbf{r}$, оставити $\mathbf{y}$ случајним, те добити номиналну рандомизациону покривеност параметра Q интервалом $\mathcal{C}$ при фиксном недостајању једнаку $1 - \alpha_f$.
- Остављамо и податке и недостајање случајним и добијамо номиналну рандомизациону покривеност параметра Q интервалом $\mathcal{C}$ при случајном недостајању једнаку $1 - \alpha_r$.

Поставља се питање тачности ових номиналних стопа, које могу бити нарушене из разних разлога, о чему смо већ раније дискутовали. Одмах за овим питањем, намеће се следеће: како да извршимо ту проверу?

Евалуација: *de jure*

Нека на располагању имамо само један реализован узорак. Теоријски гледано, за три претходна случаја процедура би била следећа:

- На основу (4.46) и коректне спецификације апостериорне расподеле параметра директно рачунамо стопу прекривања и поредимо је са номиналном;
- На основу (4.44) и коректне спецификације података при услову недостајања, рачунамо стопу покривања и поредимо са номиналном;
- На основу (4.45) и коректне спецификације механизма недостајања рачунамо стопу покривања и поредимо са номиналном.

Генерално, ми расподеле које нам требају да ово све срачунамо нећемо знати, а чак и да их знамо, најбоље чему можемо да се надамо јесте закључак да ли процедура ради добро или лоше **за те моделе**. Због тога се у пракси процедуре овако не евалуирају.

Евалуација: *de facto*

У пракси, ипак, углавном имамо коначну популацију из које вадимо узорке према неком плану узорковања, па би све стопе покривања евалуиране као у претходном под-потпоглављу биле везане строго за закључке на тој популацији. Ми бисмо ипак, да евалуирамо процедуру саму за себе, невезано од популације на којој се примењује. Стога, у пракси се раде следећи кораци:

1. Генерише се низ популација и узорака за које се очекује да ће процедура на њима бити примењивања.
2. За сваки елемент тог низа рачунају се све три вероватноће покривања - једна бајесовска $\mathbf{P}\{Q \in \mathcal{C} \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}\}$ и две рандомизационе: $\mathbf{P}\{Q \in \mathcal{C} \mid \mathbf{r}\}$ и $\mathbf{P}\{Q \in \mathcal{C}\}$.
3. Добијене резултате поредимо са номиналним стопама и бележимо потенцијална одступања.

У идеалним условима, бајесовска стопа треба да буде једнака номиналној за сваки од узорака и сваку од популација. У најгорем случају, требало би да такав буде просек. Слично важи и за остале две. Ако је номинална стопа покривања једнака просечној стварној, онда за процедуру по којој се конструише интервална оцена кажемо да је *глобално калибрисана*.

Након свега реченог, посматрајмо израз:

$$\mathbf{P}\big\{Q \in \mathcal{C} \mid \mathfrak{A}\big\}.$$

Три су нам се ствари до сада појавиле на месту $\mathfrak{A}$: $\{o(\mathbf{y},\mathbf{r}),\mathbf{r}\}$, затим $\mathbf{r}$ и коначно ништа. Посматрајмо горњу вероватноћу као функцију оног што је садржано у $\mathfrak{A}$ (шта год то било од пређашње троје). Можемо рачунати

$$\begin{aligned}\mathbf{E}\Big(\mathbf{P}\big\{Q\in\mathcal{C}\mid\mathfrak{A}\big\}\Big)&=\mathbf{E}\left[\mathbf{E}\left(I\big\{Q\in\mathcal{C}\big\}\mid\mathfrak{A}\right)\right]\\&=\mathbf{E}I\big\{Q\in\mathcal{C}\big\}\\&=\mathbf{P}\big\{Q\in\mathcal{C}\big\}.\end{aligned}$$

Приметимо да, ако на место $\mathfrak{A}$ ставимо једну од три поменуте могућности, добијамо просечне стопе покривања за сваки од видова конструкције интервала. Дакле, важи веома занимљив резултат, који смо преузели из [28], а дајемо га наредном теоремом.

Теорема 4.3. ***Ако су бајесовска, фреквенционистичка за фиксан одговор, и фреквенционистичка за случајан одговор, интервална оцена све три глобално калибрисане (за исти низ), онда морају имати исту номиналну стопу покривања.*** ■

Сада ћемо мало продискутовати појмове и резултате које смо изложили.

Калибрација: дискусија

Сходно теореми 4.3, за интервалну оцену $\mathcal{C}$ параметра Q можемо рећи да је глобално калибрисана ако је

$$1-\alpha=\mathbf{P}\big\{Q\in\mathcal{C}\big\},$$

не бринући који вид покривања је спецификован. Рећи ћемо да је интервална оцена $\mathcal{C}$ *конзервативно калибрисана* уколико јој је просечна стварна стопа покривања већа од номиналне, то јест ако је

$$\mathbf{P}\big\{Q\in\mathcal{C}\big\}>1-\alpha.$$

Згодно је уочити да глобална калибрисаност заправо и није неко превише јако својство, јер смо је дефинисали за низ могућих популација и узорака које смо генерисали од нуле. Боље би било да је интервална оцена калибрисана за све *идентификабилне* узорке и популације, то јест све оне које се могу препознати из оног „у шта гледамо", а то је $\{o(\mathbf{y},\mathbf{r}),\mathbf{r}\}$. Размотримо на примеру.

Пример 4.2. Познато је[31] да је, при простом случајном узорковању обима n из популације обима n, без присуства недостајања, 95%-ни интервал поверења за популацијску средину $\overline{Y}$ дат са

$$\overline{y}\pm\Phi^{-1}(1-0.05/2)\cdot s\cdot\sqrt{\frac{1}{n}-\frac{1}{N}},\tag{4.47}$$

где је $\overline{y}$ узорачка средина, а s узорачка стандардна девијација.

Претпоставимо сада да је сваки елемент популације, Y_i, једнак нула или 1. Онда кад је $\overline{y}=0$, биће и

$$(n-1)s^2=\sum_{j=1}^{n}(y_j-\overline{y})^2=\sum_{j=1}^{n}y_j^2=\sum_{j=1}^{n}y_j=n\overline{y}=0,$$

па је и $s=0$, те је цео интервал поверења једнак скупу $\{0\}$. Напоменимо да смо мало злоупотребили ознаке: свака сума у горњем реду иде по елементима узорка, а не по првих n из популације.

[31] [12].

Елем, било би веома погрешно рећи да $\{0\}$ има стопу покривања једнаку 95%, иако ће тај „интервал" бити приближно глобално калибрисан (неке ће популације имати више нула, неке више јединица, па ће се у просеку скратити).

Треба наћи бољу дефиницију калибрације. Генерално, рећи ћемо да је интервална оцена *апсолутно калибрисана*, ако важи да је

$$\mathbf{P}\big\{Q \in \mathcal{C} \mid o(\mathbf{y}, \mathbf{r}), \mathbf{r}\big\} = 1 - \alpha$$

за све могуће $\{o(\mathbf{y}, \mathbf{r}), \mathbf{r}\}$, а не у просеку. Ипак, за све реалне ситуације није реално очекивати да апсолутна калибрација стоји. Стога, можемо се надати да је интервална оцена *условно калибрисана* при услову h, што се дефинише као

$$\mathbf{P}\big\{Q \in \mathcal{C} \mid h\big(o(\mathbf{y}, \mathbf{r}), \mathbf{r}\big)\big\} = 1 - \alpha,$$

поново за све вредности услова. Функцију h треба бирати тако да сумира што више важних карактеристика својих аргумената. Најгрубља форма условне калибрације је глобална калибрација (где услова и нема), а најфинија форма је апсолутна калибрација где је h идентитет.

Напомена. За интервалну оцену чија је стварна стопа покривања мања од номиналне кажемо да је *либерално калибрисана*. То генерално није пожељна особина.

4.7.6 Сличност бајесовских и фреквенционистичких процедура

Смисао овог потпоглавља биће да нам да општи оквир у којем смо до сада изводили закључке, а нарочито да оправда наше често „укрштање" бајесовског и фреквенционистичког закључивања. Приликом конструкције интервалних оцена, у многим случајевима и фреквенционистичке и бајесовске оцене дају исте резултате, то јест продукују исте интервалне оцене. На пример, интервал (4.47) добија се и као 95%-ни интервал поверења при простом случајном узорковању без понављања (за велике обиме узорка), али и као 95%-ни бајесовски интервал прекривања при претпоставци апостериорне нормалности. Ми ћемо сада изложити неке елементарне резултате по овом питању, а детаљније се може прочитати у [28] и тамошњим референцама.

Нека имамо параметар $\boldsymbol{\theta}$ чија је права вредност $\boldsymbol{\theta}_0$ и нека су $Z_1, \ldots, Z_n$ IID из расподеле која зависи од $\boldsymbol{\theta}$. Нека су $\hat{\boldsymbol{\theta}}$ и T апостериорно очекивање и дисперзија (коваријациона матрица) параметра $\boldsymbol{\theta}$ при реализованом узорку $Z_1, \ldots, Z_n$. Тада при веома слабим условима регуларности (видети [32]) важи асимптотска при $n \to +\infty$ нормалност

$$[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})T^{-1/2} \mid Z_1, \ldots, Z_n] \to \mathcal{N}(\mathbf{0}, \mathbf{Id}),$$

где нам је $\mathbf{Id}$ јединична матрица.

Такође, посматрајући апостериорно очекивање $\hat{\boldsymbol{\theta}}$ као случајну функцију узорка $Z_1, \ldots, Z_n$, важе следећа два својства:

1. Апостериорно очекивање тежи нормалној расподели центрираној око праве вредности параметра:

$$[\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0] \to \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{D}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}));$$

2. Расподела апостериорне дисперзије (као функције узорка) тежи расподели која има очекивање једнако дисперзији апостериорног очекивања, а дисперзију реда величине мањег од дисперзије апостериорног очекивања:

$$[T \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0] \to (\mathbf{D}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}), \ll \mathbf{D}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}})).$$

Ознаку смо већ имали. Подсетимо се, уколико је параметар вишедимензион, горњи услов за дисперзију значи да свака компонента од T има дисперзију значајно мању од било које позитивне компоненте од $\mathbf{D}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}})$.

Последица горња два резултата је та да при правој вредности параметра $\boldsymbol{\theta}$ важи конвергенција у расподели

$$[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)T^{-1/2} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0] \to \mathcal{N}(\mathbf{0}, \mathbf{Id}).$$

Ми нећемо до у детаље улазити у то када и зашто ове апроксимације важе, јер таква тематика није директно везана за рад са недостајућим подацима, што је наша тема интересовања.

У пракси

Претходни закључци су нам практично битни, јер ми веома често, било да закључујемо фреквенционистички, било бајесовски, полазимо од нормалности разлике $\hat{Q} - Q$ са очекивањем нула и неком дисперзијом T, где при оба вида закључивања користимо исте или сличне статистике.[32] Наравно, условљавати комплетним и доступним подацима није иста ствар, али се у пракси показало да су резултати приближно исти.[33]

Штавише, није за веровати фреквенционистичком закључивању којем се не може приписати бајесовска природа, јер оно није у стању да урачуна никакво предзнање истраживача о параметру, а које се иначе даје преко априорне расподеле. Слично, није за веровати ни бајесовском закључивању које нема фреквенционистичку интерпретацију.

Значај дуалности бајесовског и фреквенционистичког приступа нарочито је дошао до изражаја у овој глави, где смо изводили вишеструко-импутационе оцене параметара.

4.8 Интервалне оцене и тестови: контекст вишеструке импутације

У претходном поглављу бавили смо се тиме на које начине се могу конструисати интервали поверења и рачунати p-вредности тестова на основу статистика које зависе од доступних података. У овом поглављу циљ ће нам бити да испричамо све специфичности које важе онда када су оцене које се користе баш вишеструко-импутационе оцене. И даље ћемо претпостављати да $\hat{Q}_l$ означава условно очекивање параметра при комплетираним подацима из l-те вишеструке импутације, то јест претпоставићемо да важе услови конгенијалности.

4.8.1 Још неке битне расподеле

Подсетимо се, при давању значаја бајесовској природи **скаларног** параметра Q и при претпоставци апостериорне нормалности, дошли смо до резултата да је

$$[Q \mid \mathcal{S}_m, B_\infty] \sim \mathcal{N}(\overline{Q}_m, \overline{U}_m + (1 + 1/m)B_\infty). \tag{4.48}$$

Уочимо да је параметар B_∞ остао неоцењен. Сада је право место да се ослободимо и тога. Како смо сличну дискусију имали и код фреквенционистичког приступа, нећемо се превише задржавати на њој, те ћемо само рећи да ће B_m, као постојана и непристрасна оцена од B_∞, због особина узорачке дисперзије у нормалном моделу који смо претпоставили, при паметном одабиру априорне расподеле[34] задовољити једнакост у расподели

$$\left[\frac{(m-1)B_m}{B_\infty} \,\middle|\, \mathcal{S}_m\right] \sim \chi^2_{m-1}. \tag{4.49}$$

Наравно, циљ је заменити B_m као оцену у (4.48). Кад би важило да је расподела од $\overline{U}_m + (1 + 1/m)B_\infty$ заправо χ^2 расподела (са ма колико степени слободе), тада би на основу познатог резултата теорије вероватноћа расподела од Q при услову само $\mathcal{S}_m$ била Студентова t расподела. Нажалост, чак иако (мало скалирано и „намештено") B_∞ има χ^2 расподелу, кад му се

[32] Подсетимо се приче о конгенијалности коју је дефинисао Менг, [25].

[33] Рубин, [28].

[34] Априорна расподела за $\ln B_\infty$ треба да буде пропорционална константи, али ми се тиме нећемо оптерећивати, јер су нам априорне расподеле и онако формалне.

дода константа $\overline{U}_m$ (константа је због услова) оно не мора задржати расподелу. Штавише, може се показати да расподела од $[Q \mid \mathcal{S}_m]$ није Студентова, већ тзв. Бехренс-Фишерова.[35]

Ипак, ова расподела може се апроксимирати t расподелом.

Апроксимација t расподелом

У ову дискусију нећемо превише улазити, већ ћемо само изложити до чега се сада дошло. Наиме, Рубин је у [28], користећи низ апроксимација за које је неопходно да важи мноштво услова извео да се може апроксимирати да је

$$[Q \mid \mathcal{S}_m] \sim t_\nu(\overline{Q}_m, T_m), \tag{4.50}$$

где је извео да је „добра" апроксимација броја степени слободе дата са

$$\nu_{\text{old}} = (m-1)\left(1 + \frac{1}{r_m}\right)^2, \tag{4.51}$$

где је r_m такозвани *релативни пораст дисперзије због постојања недостајања*,[36] тј.

$$r_m = \frac{\left(1 + \frac{1}{m}\right) B_m}{\overline{U}_m}. \tag{4.52}$$

Уочимо да је за овако дефинисан број степени слободе најнижа могућа вредност једнака $m-1$, онда када r_m „оде" у бесконачност и тада је скоро сва апостериорна дисперзија последица недостајања. Највећа могућа вредност је бесконачно и тада је сва дисперзија последица узорковања.

Ипак, ми бисмо да задржимо интерпретабилност броја степени слободе, а то је да је он једнак обиму узорка умањеном за број оцењених параметара. Барнард и Рубин су у [19] уочили да овако апроксимиран број степени слободе може узети вредност већу од обима комплетног узорка, што су описали као „очигледно неадекватно". Они су развили модификовану оцену[37] која почива на интерпретабилном доживљавању броја степени слободе које смо поменули. Нека нам је n обим узорка. Тада ћемо са ν_{com} означити број степени слободе на теоријски комплетном узорку. То се може апроксимирати тиме што се од обима узорка одузме број параметара које модел има. Уколико дефинишемо *удео дисперзије која се може приписати недостајању*:

$$\lambda_m = \frac{\left(1 + \frac{1}{m}\right) B_m}{T_m}, \tag{4.53}$$

онда дефинишемо број степени слободе *за доступне податке* као

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3}\nu_{\text{com}}(1 - \lambda). \tag{4.54}$$

Након што су увели све ове појмове, Барнард и Рубин предложили су модификовану оцену броја степени слободе дату са

$$\nu = \frac{\nu_{\text{old}}\nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}}. \tag{4.55}$$

Многе досадашње апроксимације деловале су потпуно насумично, па је ред да објаснимо откуд оне. Наиме, овако дефинисано ν увек је мање или једнако од ν_{com}, што је свакако пожељно. Затим, ако је $\nu_{\text{com}} = +\infty$ (у лимесу), онда је $\nu = \nu_{\text{old}}$. Ако је $\lambda = 0$, тј. ако нема недостајања, онда се своди да је $\nu = \nu_{\text{com}}$, а ако је $\lambda = 1$, сва дисперзија је услед недостајања, налазимо $\nu = 0$. Расподеле са нула степени слободе нису дефинисане, што овде тумачимо као да нам доступни подаци не дају никакву информацију о параметру, па нема смисла ни спроводити никакву статистичку анализу.

[35] Видети [28], страна 90.

[36] Зове се тако јер представља количник оцене дисперзије оцене по недостајању при фиксним доступним подацима, то јест оцене од B_∞ и оцене „обичне" дисперзије, без недостајања, $\overline{U}_\infty$.

[37] Отуда и ознака „old" (стара) у претходној апроксимацији.

Постоје радови који предлажу друге оцене броја степени слободе, као и радови који пореде све те оцене, а о њима се може прочитати у [6], страна 48.[38]

За скаларни параметар Q се сада на основу овако апроксимиране расподеле формирају интервали поверења, тестирају хипотезе и слично. Штавише, због дуалности бајесовског и фреквенционистичког приступа расподела се може задржати и за френквенционистичку природу $\hat{Q}$ и $\hat{U}$, али ми у те специфичности нећемо улазити.[39]

4.8.2 Вишеструко-импутационе p-вредности за коначно m и вишедимензиони параметар

У пододељку 4.4.4 видели смо како се изводи апостериорна расподела при услову $\mathcal{S}_m$ и B_∞ у случају скаларног параметра. За k-димензионално Q процедура је истоветна све до једног тренутка, а тај тренутак јесте када треба оценити B_∞ помоћу B_m. Овде ћемо покушати то да решимо.

Јасно, вишедимензионалност параметра неће „покварити" резултат

$$[Q \mid \mathcal{S}_m, B_\infty] \sim \mathcal{N}(\overline{Q}_m, \overline{U}_m + (1 + 1/m)B_\infty). \tag{4.56}$$

Уколико бисмо знали B_∞, и уколико бисмо хтели да тестирамо да ли је $Q = Q_0$, могли бисмо се послужити једнакошћу у расподели

$$(Q_0 - \overline{Q}_m)[\overline{U}_m + (1 + 1/m)B_\infty]^{-1}(Q_0 - \overline{Q}_m)^T \sim \chi^2_k,$$

те добити p-вредност нулте вредности Q_0 као

$$\textbf{p-val}(Q_0 \mid \mathcal{S}_m, B_\infty) = \mathbf{P}\left\{\chi^2_k > (Q_0 - \overline{Q}_m)[\overline{U}_m + (1 + 1/m)B_\infty]^{-1}(Q_0 - \overline{Q}_m)^T\right\}. \tag{4.57}$$

Ипак, ако не знамо B_∞, онда је коректна бајесовска p-вредност нулте вредности Q_0 дата на основу формуле потпуне вероватноће:

$$\textbf{p-val}(Q_0 \mid \mathcal{S}_m) = \int \mathbf{P}\left\{\chi^2_k > (Q_0 - \overline{Q}_m)[\overline{U}_m + (1 + 1/m)B_\infty]^{-1}(Q_0 - \overline{Q}_m)^T\right\} f(B_\infty \mid \mathcal{S}_m)\,\mathrm{d}B_\infty. \tag{4.58}$$

Овде се јављају два проблема: први, како наћи расподелу од B_∞ при услову S_m и други, како срачунати горњи интеграл. Први проблем се решава лакше: применом бајесове формуле добијемо да је жељену расподелу могуће записати као производ расподеле B_m при услову B_∞, за коју знамо да је Висхартова (имали смо у одељку 4.6), и априорне расподеле за B_∞. Ипак, рачунање горњег интеграла није ни најмање једноставно.

Ad hoc решење које је предложио Рубин 1987. јесте да се у тест статистици

$$(Q_0 - \overline{Q}_m)[\overline{U}_m + (1 + 1/m)B_\infty]^{-1}(Q_0 - \overline{Q}_m)^T$$

B_∞ просто замени са B_m, и да се користи тест статистика

$$D_m = (Q_0 - \overline{Q}_m)[\overline{U}_m + (1 + 1/m)B_m]^{-1}(Q_0 - \overline{Q}_m)^T \tag{4.59}$$

чија ће се расподела апроксимирати Фишеровом $F_{k,\nu_{\text{old}}}$. Ова оцена је, према самом Рубину, јако лоша и коришћена расподела је слабо оправдана.

Постоје разне финије апроксимације: при скаларности параметра, при априорној сразмерности параметара B_∞ и T_∞ итд. Ми се њима нећемо бавити детаљно, а све се могу наћи у [28], поглавље 3.4.

[38] Пратећи, наравно, тамошње референце.

[39] Прође се кроз сличну дискусију као у пододељку 4.5.3.

4.8.3 Комбиноване *p*-вредности

У пракси, ипак, већина статистичких алата неће вратити оцену коваријационе матрице, већ ће просто, на основу комплетних/комплетираних података вратити реализовану вредност тест статистике и одговарајућу p-вредност. Стога општу p-вредност након m допуњавања података треба тражити само на основу вредности

$$d_l = (Q_0 - \hat{Q}_l)\hat{U}_l^{-1}(Q_0 - \hat{Q}_l)^T, \tag{4.60}$$

и одговарајућих p-вредности

$$\mathbf{p\text{-}val}_l = \mathbf{P}\{\chi_k^2 > d_l\}. \tag{4.61}$$

За k-димензионално Q Рубин је предложио да се реализована вредност тест статистике апроксимира са

$$\hat{D}_m = \frac{\frac{\overline{d}_m}{k} - \frac{m-1}{m+1}r_m}{1 + r_m} \tag{4.62}$$

и да се користи референтна $F_{k,(k+1)\nu_{\text{old}}/2}$ расподела. Оправдања су једва и хеуристичка, и то за случај поменуте априорне сразмерности.

У литератури нисмо успели да нађемо општа побољшања Рубинових решења, осим у [26], где сам Рубин овакву апроксимацију описује речима: „далеко од оног чему се можда надамо".[40]

Ово питање, дакле, умногоме остаје отворен проблем.

4.9 Неки примери метода вишеструке импутације

Након што смо теоријски образложили коришћење вишеструке импутације (онолико колико је то данашња статистика у стању), време је и да наведемо неколико конкретних метода и алгоритама којима се она врши. Нећемо се бавити давањем конкретних алгоритама, у строгом смислу „корак по корак", ових метода, али заинтересовани их, скоро све, могу пронаћи у [6].

4.9.1 Недостајање само у једној колони

У овом пододељку навешћемо неколико примера вишеструке импутације онда када је само једна колона захваћена недостајућим подацима. Оваква појава честа је у истраживањима у којима се примењује линеарна регресија, а недостајања углавном постоје само у циљној променљивој.

Пример 4.3 (*predictive Mean Matching*)**.** Са овим алгоритмом сусрели смо се у контексту једноструке импутације и то у примеру 3.6. Тамо смо описали како алгоритам ради и које су му предности, а овде ћемо описати како се он модификује за рад са вишеструком импутацијом.

Заправо, то ће бити лак посао, јер смо ми импутирали тако што смо из скупа блиских комплетних инстанци[41] случајним одабиром бирали једну, чије смо одговарајуће поље користили за „донора" и њиме импутирали. Код вишеструке импутације, просто, тај одабир вршићемо m пута.

Предности су све оне које смо навели у примеру 3.6, док је у контексту вишеструке импутације очигледна мана често понављање вредности којима импутирамо.

Слично се за потребе вишеструке импутације прилагођавају и остали методи из хот дек фамилије.

Пример 4.4 (CART)**.** Алгоритми из групе класификационих и регресионих дрвета одлучивања[42] честа су алтернатива линеарној и логистичкој регресији, због своје неосетљивости на претпоставке модела, или, боље речено, због тога што се не тражи да се икакав модел уопште

[40] Енг. *far from what might be hoped for.*

[41] У смислу у којем смо дефинисали блискост у примеру 3.6.

[42] Енг. *classification and regression trees - CART.*

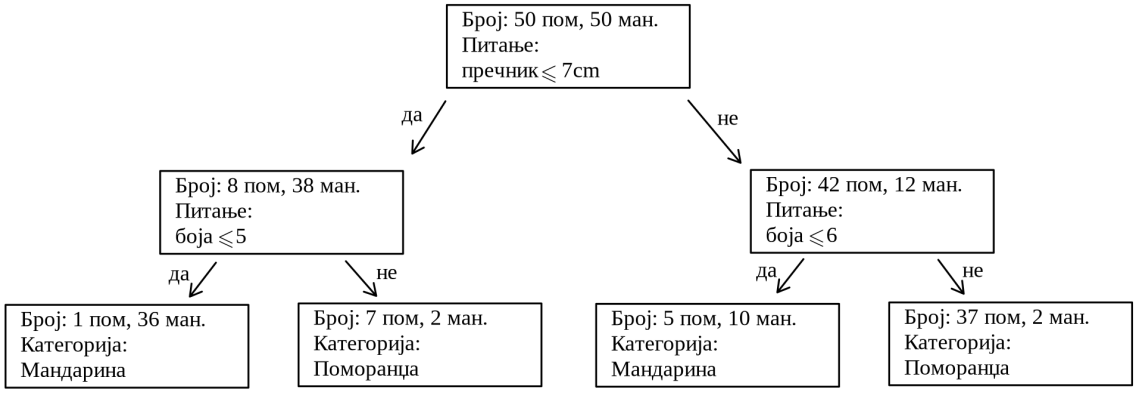
и спецификује. За почетак описаћемо како CART бива имплементиран за потребе класичне регресије и класификације, па ћемо после у причу укључити и недостајуће податке.

Дрвета одлучивања функционишу тако што се крене од корена дрвета, у којем се поставља питање везано за неки атрибут у подацима. Рецимо, за нумеричке атрибуте често је питање „да ли је вредност атрибута на инстанци мања или већа од неке задате вредности?"; тим питањем се скуп инстанци раздваја на две категорије, од којих се свака раздваја на нове две, новим питањем. Дрвета одлучивања јесу алгоритам *надгледаног* учења, то јесте захтева се лабелираност података и оно што алгоритам учи јесте *оптимално дрво*.

Узмимо, на пример,[43] проблем бинарне класификације, у којем је циљ научити оптимално дрво које раздваја поморанџе од мандарина. Нека су нам атрибути следећи:

- Пречник у центиметрима, који је корисно посматрати јер су поморанџе у просеку веће од мандарина.
- Боја, на скали која почиње од 0, где већи број представља тамнију нијансу наранџасте боје. Претпоставимо тако да су поморанџе на овој скали дате бројем 5 или мање. Ово је корисно посматрати јер су поморанџе у просеку светлије од мандарина.

Ми, дакле, на основу скупа који нам је дат, знамо шта су поморанџе, а шта мандарине. Једно могуће дрво има облик као на слици 4.1.



Слика 4.1: Пример дрвета одлучивања за бинарну класификацију

Последњи (доњи) чворови у дрвету зову се листови, и уколико нека нова инстанца прошавши кроз скуп питања из дрвета заврши у неком листу, класификује се у ону категорију која је у том листу била заступљенија при изградњи дрвета на тест скупу. У случају регресије, инстанци се додељује вредност циљне променљиве једнака просеку циљне променљиве јединки из тест скупа које се налазе у том листу.

Мењајући питања, као и њихов распоред, дрво се оптимизује. Мере квалитета су стандардне: средњеквадратна грешка за регресију и прецизност (тачност и одзив) за класификацију. Јасно је да у корен дрвета треба ставити питање које добро раздваја инстанце. Инстанце ће бити добро раздвојене неким питањем уколико је хомогеност унутар сваке подгрупе велика, што се може квантификовати *ентропијом* или *Џинијевим индексом*. Ми их нећемо детаљније описивати, а о њима се детаљније може прочитати у [2].

Прилагодба за импутацију, барем једноструку, је очигледна. На комплетним инстанцама формира се дрво, кроз које се онда „пропусте" предиктори за дату некомплетну инстанцу, те се предвиђена вредност циљне променљиве користи као импутациона. За потребе вишеструке импутације процедура је следећа: након што инстанца „стигне" у одговарајући лист, случајно се бира тестна јединка из тог листа, те се њена вредност циљне променљиве користи као импутациона. Овиме се урачунава неодређеност о недостајућој вредности.

Наравно, као и код PMM алгоритма, нерепрезентативност узорка може да се компензује тако што се при свакој наредној импутацији формира ново дрво, и то на основу новог бутстреп

[43] Пример смо преузели из [34].

узорка из скупа комплетних инстанци. О разним другим модификацијама, као и рачунарској имплементацији може се прочитати у [6].

Напоменимо и да се дрвета одлучивања у машинском учењу углавном не користе самостално, већ се гради већи број дрвета чија се предвиђања комбинују у један модел, тзв. *ансамбл*. О овоме се може информисати из [2].

4.9.2 Недостајања у више колона

У претходном потпоглављу дали смо пар примера алгоритама за импутацију недостајућих података онда када недостајања постоје само у једној колони у подацима. Још у примеру 3.6 видели смо како се један конкретан алгоритам (PMM) може модификовати да импутира и вишедимензионо недостајање; слично се може испричати за већину метода. Ипак, ми ћемо сада испричати мало општију причу, која неће зависити од конкретне имплементације конкретног метода или алгоритма. Користићемо ознаку Y_{-j} за податке из којих је избачена j-та колона.

Приметимо, у позадини сваког метода импутације, било једноструке или вишеструке, стоји искоришћавање везе међу колонама: што боље искористимо везу колоне у коју уписујемо вредност и осталих колона, то ће уписана вредност бити квалитетнија. Међутим, када су недостајања насумично „разбацана" по подацима, искоришћавање ове везе постаје знатно теже. Бурен у [6] идентификује основне проблеме који настају при вишедимензионом недостајању:

- Саме колоне Y_{-j} могу да садрже недостајуће податке, што онемогућава њихово коришћење за предвиђање;
- Може доћи до „кружне" зависности - недостајући подаци у једној колони могу да зависе од недостајућих података у другој колони, и обрнуто;
- Подаци у колонама су углавном различитог типа (нумерички, категорички итд) што онемогућава уклапање података у неку класичну расподелу вероватноћа, рецимо вишедимензиону нормалну итд.

Дугогодишњом применом разних метода, временом су се искристалисала три генерална приступа за импутацију при вишедимензионом недостајању, а то су:

1. **Импутација монотоног недостајања**. За монотоне обрасце недостајања (које смо дефинисали на самом почетку), импутира се узастопном применом метода импутације за једнодимензионо недостајање.
2. **Заједничко моделовање.** За нестандардне обрасце недостајања претпоставља се расподела за податке и импутационе вредности се узоркују из ње.
3. **Пуна условна спецификација**[44], познатија као **уланчане једначине** или **секвенцијална регресија**. Модел за вишедимензионо недостајање је спецификован преко низа условних модела за једнодимензионо и импутационе вредности се узоркују из итерираних условних модела.

Монотоно недостајање

Претпоставимо да су променљиве у подацима дате са $Y_1, \ldots, Y_K$ и да су поређане тако да недостајање у подацима чини монотон образац, као на слици 1.1. Импутира се на следећи начин. Потенцијалне недостајуће вредности у колони Y_1 импутиране су било на основу претходних сазнања, било на неки други начин. Често међу колонама постоји бар једна комплетна, па се она може употребити као предиктор за импутацију у Y_1. Затим, недостајуће вредности у Y_2 бивају импутиране коришћењем (сада комплетне) колоне Y_1 и тако, до краја, када Y_K бива импутирано коришћењем колона $Y_1, \ldots, Y_{K-1}$ као предиктора.

Било који метод који се користи за импутацију једнодимензионог недостајања може бити искоришћен као „градивни елемент" за импутацију монотоног недостајања. За вишеструку

[44] Енг. *fully conditional specification.*

импутацију цео процес се понавља m пута (наравно, макар почетни корак мора укључивати неки случајни одабир). Више о овом недостајању и његовој импутацији може се прочитати у [6] и [28].

Заједничко моделовање

Код овог приступа полазимо од претпоставке да се подаци могу описати вишедимензионом расподелом, па се, игноришући недостајање, импутационе вредности узоркују из те расподеле. У пракси, нека инстанца може имати једно недостајуће поље, нека два и слично, па се онда расподела недостајућих података, при услову доступних, мења од врсте до врсте. Најчешће коришћена расподела је, наравно, вишедимензионална нормална.

Јасно је да ће параметри модела којим импутирамо, назовимо их $\boldsymbol{\zeta}$, зависити од параметара расподеле података $\boldsymbol{\theta}$, које такође не знамо, а углавном их не можемо оценити директно из доступних података.

Када су подаци нумерички, а претпостављена расподела података нормална, користи се тзв. *аугментација података*[45] која, начелно говорећи, смењује кораке импутације и оцењивања параметара $\boldsymbol{\theta}$, до конвергенције. Детаљан опис алгоритма налази се у [6], страна 116. За категоричке податке нема универзалног приступа, а и о њима може се читати у [6]. Вишеструка импутација се, наравно, добија вишеструким узорковањем.

Пуна условна спецификација

Пуна условна спецификација, коју ћемо надаље ословљавати енглеском скраћеницом FCS, врши импутацију тако што за сваку колону Y_j формира расподелу вероватноћа $f(\mathbf{Y}, \mathbf{R})$ кроз скуп условних расподела $f(Y_j \mid Y_{-j}, R)$. Почне се од насумичног попуњавања, те се кроз **све** међусобне условне моделе итерира M пута до жељене границе. Рубин је[46] у процесу FCS издвојио три корака: корак *моделовања*, где бирамо малопре поменути модел, затим корак *оцењивања*, где се при том моделу формирају апостериорне расподеле/оцене параметара, те на крају корак *импутације* који узоркује из формираних расподела. FCS проблем спецификовања вишедимензионе расподеле за податке своди на спецификовање једнодимензионих условних.

Пример 4.5 (MICE)**.** Један од најкоришћенијих алгоритама јесте тзв. MICE алгоритам, што представља скраћеницу од енглеског *Multivariate Imputation by Chained Equations*, односно *вишеструка импутација путем уланчаних једначина*. Овај алгоритам конструисали су Бурен и Гротиус-Оудсхорн у [17]. Договоримо се да са R_j означимо j-ту колону реализованог обрасца недостајања (овде га гледамо као матрицу). Тада је алгоритам MICE дат корацима:

1. Спецификовати модел $f(\overline{o}(Y_j, R_j) \mid o(Y_j, R_j), Y_{-j}, R)$ за свако $j = 1, \ldots, K$.
2. За свако j, направити иницијалне импутације $\dot{Y}_j^0$ случајно узоркујући из $o(Y_j, R_j)$.
3. Почети итерирање за $t = 1, \ldots, M$.
4. Почети итерирање за $j = 1, \ldots, K$.
5. Дефинисати $\dot{Y}_{-j}^t$ као тренутно комплетиране податке без j-те колоне.
6. При услову $\dot{Y}_{-j}^t$ и $o(Y_j, R_j)$ узорковати параметре модела из тачке 1. из њихове апостериорне расподеле.
7. Из расподеле у 1, при узоркованим параметрима из претходног корака, узорковати $\dot{Y}_j^t$.
8. Завршити итерирање по j.
9. Завршити итерирање по t.

[45] Енг. *data augmentation*.
[46] Рубин, [28].

Уочимо да при оваквом алгоритму имамо задате само условне расподеле, а не и заједнучку расподелу. Уколико уочимо да је процес који MICE генерише један ланац Маркова, са скупом стања једнаком скупу свих могућих импутационих вредности, можемо се запитати о конвергенцији. По овом питању још увек се врше истраживања (када ће процес исконвергирати правој расподели), а о томе докле се стигло може се прочитати у [6] и тамошњим референцама.

Глава 5

Тестирање вишедимензионе нормалности

5.1 Поставка проблема

Претпоставимо да имамо прост случајан узорак $\mathbf{X}_1, \ldots, \mathbf{X}_n$ d-димензионих случајних вектора, од којих сваки има расподелу вероватноћа $\mathbf{P}^{\mathbf{X}}$. Уколико означимо да нам је

$$\mathcal{N}_d := \left\{\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \text{ позитивно дефинитна}\right\}$$

скуп свих недегенерисаних d-димензионих нормалних расподела, ми заправо желимо да тестирамо да ли је

$$\mathbf{P}^{\mathbf{X}} \in \mathcal{N}_d,$$

то јест да ли је наш узорак из вишедимензионе нормалне расподеле, против уопштених алтернатива (дакле не против једне конкретне алтернативне расподеле). Ми желимо да направимо тест са тест статистиком T_n која ће дати адекватан одговор на ово питање.

Напомена. Уколико не нагласимо супротно, сваки вектор сматраћемо колона-вектором.

Тест статистика T_n би, наравно, требало да задовољава одређене особине. Поред класичних особина које се од квалитетног статистичког теста очекују, као што су непристрасност, постојаност итд, код тестирања вишедимензионе нормалности разумно је очекивати да важи и тзв. *афина инваријантност*, то јест да је

$$T_n(\mathbf{A}\mathbf{X}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{X}_n + \mathbf{b}) = T_n(\mathbf{X}_1, \ldots, \mathbf{X}_n),$$

кад год је $\mathbf{A}$ d-димензиона инвертибилна матрица, и $\mathbf{b} \in \mathbb{R}^d$ произвољно. Заиста, ово има смисла тражити: уколико податке који су нормално расподељени мало ротирамо и транслирамо, они ће остати нормално расподељени, па нема смисла да пре ротације, рецимо, прихватимо H_0, а након ротације одбацимо.

Уколико претпоставимо да је наш узорак из апсолутно непрекидне расподеле вероватноћа (у односу на Лебегову меру), тада ће узорачка коваријациона матрица,[1] $\mathbf{S}_n = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}}_n)(\mathbf{X}_j - \overline{\mathbf{X}}_n)^T$, скоро сигурно бити регуларна ([4]), па има инверз $\mathbf{S}_n^{-1}$. Како је ова матрица симетрична и регуларна, она има јединствен корен $\mathbf{S}_n^{-1/2}$. Сада за свако $j = 1, \ldots, n$ можемо посматрати

$$\mathbf{Y}_{n,j} = \mathbf{S}_n^{-1/2}(\mathbf{X}_j - \overline{\mathbf{X}}_n).$$

Идеја овакве трансформације јесте да од података направимо податке са нултом средњом вредношћу и јединичном коверијационом матрицом. Овакви трансформисани подаци свој пуни смисао добиће у наредном одељку.

[1] Која се некада дефинише делећи са $n-1$, а не са n, како би се добила непристрасна оцена. Ипак, оваква модификација не утиче на асимптотске одлике оцена, те се њоме нећемо превише оптерећивати.

Пожељне особине тест статистике

Нагласићемо још и да је за афину инваријантност довољно да тест статистика T_n зависи искључиво од $\mathbf{Y}_{n,1},\dots,\mathbf{Y}_{n,n}$ и да за сваку ортогоналну матрицу $\mathbf{O}$ реда d важи да је $T_n(\mathbf{OY}_{n,1},\dots,\mathbf{OY}_{n,n}) = T_n(\mathbf{Y}_{n,1},\dots,\mathbf{Y}_{n,n})$. Такође, довољно за афину инваријантност је и да T_n зависи само од израза $\mathbf{Y}_{n,i}^T\mathbf{Y}_{n,j}$. И о овоме се може пронаћи у [4], пратећи тамошње референце.

Уколико је статистика T_n афино инваријантна, онда њена расподела под нултом хипотезом не зависи од $\boldsymbol{\mu}$ и $\boldsymbol{\Sigma}$, па се без умањења општости може претпоставити да је нулта расподела d-димензиона стандардна нормална. Штавише, уколико је статистика T_n афино инваријантна, чак и под било којом алтернативом за коју је $\mathbf{E}\|\mathbf{X}\|^2 < +\infty$, нема сметње претпоставити да је $\mathbf{EX} = \mathbf{0}$ и $\mathbf{EXX}^T = \mathbf{Id}$.

Напомена. Где год не нагласимо на коју се норму мисли, мисли се на еуклидску норму.

Испитивање особина (попут, рецимо, постојаности) великог броја статистичких тестова најчешће се врши емпиријски, симулацијама. Ипак, то скоро никада није крајњи циљ истраживача: симулациона студија углавном представља добар водич ка томе шта треба доказати теоријски. Због тога, тест статистика T_n треба да буде „смислена". То значи да би T_n, евентуално нако неке трасформације, требало да представља оцену неког инваријантног функционала $\mathcal{T}(\mathbf{P}^{\mathbf{X}})$, где под инваријантношћу подразумевамо то да уколико је $\tilde{\mathbf{X}}$ регуларна афина слика од $\tilde{\mathbf{X}}$, онда је и $\mathcal{T}(\mathbf{P}^{\tilde{\mathbf{X}}}) = \mathcal{T}(\mathbf{P}^{\mathbf{X}})$. Специјално, $\mathcal{T}$ треба да буде константан на класи $\mathcal{N}_d$.

Да би се могла очекивати постојаност теста заснованог на статистици T_n, за коју очекујемо да барем у вероватноћи тежи ка правој вредности функционала ког оцењује, функционал мора карактерисати нулту хипотезу. Прецизније, уколико је расподела вероватноћа $\mathbf{P}^{\mathbf{X}_1}$ из нулте хипозете, а $\mathbf{P}^{\mathbf{X}_2}$ није, онда мора важити да је

$$\mathcal{T}(\mathbf{P}^{\mathbf{X}_1}) \neq \mathcal{T}(\mathbf{P}^{\mathbf{X}_2}).$$

5.2 Статистике засноване на тежинским L^2 растојањима

Статистике засноване на тежинским L^2 растојањима (енг. *Weighted L^2 Statistics*) за тестирање H_0 јесу све оне које су облика[2]

$$T_n(\mathbf{X}_1,\dots,\mathbf{X}_n) = \int Z_n^2(\mathbf{t})w(\mathbf{t})\mathrm{d}\mathbf{t},$$

где нам је $Z_n(\mathbf{t}) = z_n(X_1,\dots,X_n,t)$ реална мерљива функција дефинисана на $(\mathbb{R}^d)^{n+1}$, а $w:\mathbb{R}^d\to\mathbb{R}$ је ненегативна тежинска функција која задовољава

$$\left(\forall(\mathbf{x}_1,\dots,\mathbf{x}_n)\in(\mathbb{R}^d)^n\right)\ \int z_n^2(\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{t})w(\mathbf{t})\mathrm{d}\mathbf{t} < +\infty. \tag{5.1}$$

Само Z_n углавном је облика

$$Z_n(\mathbf{t}) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} l(\mathbf{t}^T\mathbf{Y}_{n,j}),\ \ \mathbf{t}\in\mathbb{R}^d, \tag{5.2}$$

где је l нека мерљива функција која задовољава да је $\int \mathbf{E}l^2(\mathbf{t}^T\mathbf{X})w(\mathbf{t})\mathrm{d}\mathbf{t} < +\infty$, $\mathbf{E}l(\mathbf{t}^T\mathbf{X}) = 0$, при нултој хипотези.

Напомена. Функција z_n може бити и векторска, а тада у изразу (5.1) под интегралом стоји квадрат њене норме.

Ова класа тест статистика добила је име по томе што је природно окружење за извођење њихових асимптотских својстава један Хилбертов простор

$$\mathbb{H} = L^2(\mathbb{R}^d,\mathcal{B}^d,w(\mathbf{t})\mathrm{d}\mathbf{t}).$$

[2]Убудуће подразумевамо да су сви неозначени интеграли преко $\mathbb{R}^d$.

Уколико са $\|\cdot\|_{\mathbb{H}}$ означимо норму у овом простору, тада је, за фиксан узорак, заправо

$$T_n = \|Z_n\|_{\mathbb{H}}^2.$$

Уочимо да $\mathbf{Y}_{n,j}$-ови у општем случају нису независни, па сума која фигурише у (5.2) није сума IID величина, а за такве суме асимптотска својства гарантује Централна гранична теорема. Стога је класична техника за испитивање асимптотских својстава T_n та да се Z_n дефинисано као у (5.2) апроксимира одговарајућим елементом $Z_{n,0}$ из $\mathbb{H}$ који је облика

$$Z_{n,0}(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l_0(\mathbf{t}^T X_j), \tag{5.3}$$

где l_0 задовољава исте особине као l и још $\|Z_n - Z_{n,0}\|_{\mathbb{H}} \xrightarrow{\mathbf{P}} 0$. Овиме се сума величина које нису IID апроксимира сумом величина које то јесу, а за такве много боље познајемо асимптотику.

5.3 ВНЕР тест

Тест *Baringhaus-Henze-Epps-Pulley*-ја представља један пример теста који је припадник класе тестова вишедимензионе нормалности о којој смо говорили у претходном поглављу. Његова тест статистика дата је са

$$\text{ВНЕР}_{n,\beta} = n \int |\Psi_n(\mathbf{t}) - \Psi_0(\mathbf{t})|^2 w_\beta(\mathbf{t}) \mathrm{d}\mathbf{t}, \tag{5.4}$$

где је

$$\Psi_n(\mathbf{t}) = \frac{1}{n} \sum_{j=1}^{n} \exp(i\mathbf{t}^T \mathbf{Y}_{n,j})$$

емпиријска карактеристична фукнција од $\mathbf{Y}_{n,1}, \ldots, \mathbf{Y}_{n,n}$, а тежинска функција $w_\beta(\mathbf{t})$ дата је са

$$w_\beta(\mathbf{t}) = (2\pi\beta^2)^{-d/2} \exp\left(-\frac{\|\mathbf{t}\|^2}{2\beta^2}\right).$$

Сматрамо да је $\beta > 0$ фиксирана константа. Такође, претпостављамо да је $\Psi_0(\mathbf{t}) = \exp\left(-\|\mathbf{t}\|^2/2\right)$ карактеристична функција нормалне $\mathcal{N}(0, \mathbf{Id})$ расподеле.

Напомена. У наставку ћемо се реферисати на мноштво познатих резултата у вези са ВНЕР тестом. Како не бисмо оптерећивали запис мноштвом референци, има се сматрати да позивање на сваки познат резултат може бити оправдано радом [4]. Тај рад представља својеврстан преглед тестова заснованих на L^2 тежинским растојањима, закључно са 2020. годином, те су резултати које ћемо изнети доказани у радовима на које се тамо реферише.

Наиме, јасно је да мале вредности тест статистике $\text{ВНЕР}_{n,\beta}$ указују на тачност нулте хипотезе. Квалитет овог теста огледа се, између осталог, у томе што је доказано да скоро сигурно важи да је

$$\liminf_{n\to+\infty} \frac{1}{n}\text{ВНЕР}_{n,\beta} \geqslant C(\mathbf{P}^{\mathbf{X}}, \beta) > 0,$$

где је $C(\mathbf{P}^{\mathbf{X}}, \beta)$ нека константа, за било које $\mathbf{P}^{\mathbf{X}}$ које није из нултог скупа, то јест не припада класи $\mathcal{N}_d$. Као последица, тест је постојан против сваке алтернативе.

Претпоставимо, без умањења општости,[3] да је $\mathbf{E}\|\mathbf{X}\|^2 < +\infty$ и да је $\mathbf{E}\mathbf{X}\mathbf{X}^T = \mathbf{Id}$. Показано је да тада важи

$$\frac{1}{n}\text{ВНЕР}_{n,\beta} \xrightarrow{\text{с.с.}} \Delta_\beta := \int |\Psi(\mathbf{t}) - \Psi_0(\mathbf{t})|^2 w_\beta(\mathbf{t}) \mathrm{d}\mathbf{t},$$

где је $\Psi(\mathbf{t}) = \mathbf{E}\exp(i\mathbf{t}^T\mathbf{X})$, $\mathbf{t} \in \mathbb{R}^d$, „права" карактеристична функција од $\mathbf{X}$, а ознака „с.с." означава скоро сигурну конвергенцију.

[3] Имали смо дискусију.

Напомена. Можемо уочити да је код овог теста функционал $\mathcal{T}$, о којем смо говорили у првом поглављу, заправо Δ_β.

Показано је да за сваку алтернативу за коју је $\mathbf{E}\|\mathbf{X}\|^4 < +\infty$ важи

$$\sqrt{n}\left(\frac{1}{n}\text{BHEP}_{n,\beta} - \Delta_\beta\right) \xrightarrow{D} \mathcal{N}(0, \sigma_\beta^2), \tag{5.5}$$

где σ_β^2 зависно од β. Нађена је и постојана оцена за σ_β^2, на основу које се може градити асимптотски интервал поверења за Δ_β.

Показано је и да $\text{BHEP}_{n,\beta}$ заиста јесте статистика заснована на тежинском L^2 растојању, и то

$$\text{BHEP}_{n,\beta} = \int Z_n^2(\mathbf{t}) w_\beta(\mathbf{t}) \mathrm{d}\mathbf{t},$$

где је

$$Z_n(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\cos(\mathbf{t}^T \mathbf{Y}_{n,j}) + \sin(\mathbf{t}^T \mathbf{Y}_{n,j}) - \Psi_0(\mathbf{t})\right).$$

Показано је, штавише, без услова $\mathbf{E}\|\mathbf{X}\|^4 < +\infty$, да

$$\text{BHEP}_{n,\beta} \xrightarrow{D} \int Z^2(\mathbf{t}) w_\beta(\mathbf{t}) \mathrm{d}\mathbf{t},$$

где је Z неки центрирани Гаусов процес. Зна се и да је тест у стању да детектује (одбаци) низ блиских алтернатива које се нултом скупу приближавају брзином мањом од $1/\sqrt{n}$.[4]

На основу претходних резултата оправдано стичемо утисак да је BHEP тест квалитетан по већини класичних мерила, те га је занимљиво даље испитивати.

5.4 Симулациона студија: поставка

Циљ наше симулационе студије биће да емпиријски испитамо како се резултати тестирања BHEP тестом мењају у зависности од тога који импутациони модел/алгоритам је коришћен. У ту свхру ми ћемо генерисати (комплетне) податке, како из нормалне расподеле, тако и из различитих алтернатива, а затим генерисати недостајања у тим подацима. Након тога, импутираћемо различитим методима и упоредити резултате теста.

У претходној глави видели смо да још увек нису познати јасни и теоријски поткрепљени начини за комбиновање p-вредности добијених из m пута „допуњених" података. Стога смо се у нашој студији одчучили за примену искључиво једноструке импутације.

5.4.1 Мера и моћ теста

Сматрамо да је разумно претпоставити да је читалац упознат са основним концептима статистичког тестирања. Подсетимо се, статистички тест врши се тако што се посматра тест статистика T и критична област $\mathcal{W}$, те уколико реализована вредност тест статистике (њена вредност у реализованом узорку) припада критичној области - одбацујемо нулту хипотезу, а иначе је прихватамо. Границе критичне области одређују се тако да при тачној H_0 дозвољавамо извесну толеранцију да се догоди *грешка прве врсте*, то јест да тачну нулту хипотезу прогласимо нетачном. Математички, ова толеранција задаје се вероватноћом грешке прве врсте:

$$\mathbf{P}_{H_0}\{T \in \mathcal{W}\}.$$

Нулта хипотеза често може бити вишечлана; на пример $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, где је $\boldsymbol{\theta}$ неки параметар од интереса, а $\boldsymbol{\Theta}_0$ неки вишечлан подскуп скупа свих могућих вредности параметра. Стога се посматра и *мера теста*, дата као

[4] У смислу да је за нулту расподелу $\mathcal{N}$ и фиксну алтернативну $\mathcal{M}$ низ алтернатива задат са $\mathcal{N} + \frac{1}{\sqrt{n}}\mathcal{M}$.

$$\sup_{H_0} \mathbf{P}_{H_0}\{T \in \mathcal{W}\}.$$

Често се обе вредности означавају са α, што ћемо и ми користити, а из контекста ће бити јасно на шта се мисли.

Уколико је нулта хипотеза тачна, очекујемо да грешку прве врсте правимо у $100\alpha\%$ случајева. Ипак, уколико из података који су заиста нормално расподељени уклонимо неке податке, па после импутирамо, та импутације може у великој мери „покварити" расподелу података, те ће тест пријавити „лажну" меру. О овоме ћемо у наставку рећи мало касније, те дати ad hoc решење које даје солидне резултате.

Јасно је да желимо тест који тачне нулте хипотезе проглашава тачним. Ипак, може се посматрати и сродно својство, а то је способност теста да „детектује" алтернативе, то јест да прогласи нетачном нулту хипотезу онда када она заиста јесте нетачна. У циљу проучавања овог својства посматра се *функција моћи теста*, која за дату алтернативу H_1 рачуна вероватноћу да је, при тој алтернативи, нулта хипотеза одбачена:

$$\mathbf{M}(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}\{T \in \mathcal{W}\},$$

где је $\boldsymbol{\theta}$ параметар који одговара алтернативи H_1.

У пракси, моћ теста се рачуна емпиријски, и то као

$$1 - \hat{F}_n(c),$$

где је $\hat{F}_n$ емпиријска функција расподеле узорка, а c критична вредност теста, то јест она вредност која дели критичну област од свог комплемента. У начелу, критична област може бити различитих облика, па би се моћ емпиријски рачунала другачије, али је нама јасно да тест који смо одабрали да посматрамо одбацује нулту хипотезу за велике вредности тест статистике (јер BHEP статистика представља извесно растојање), па је оваква поставка разумна.

5.5 Симулациона студија

Напомена. Сви резултати симулација које будемо дали у наставку срачунати су за обим појединачног узорка једнак 50 и за 10000 симулација, јер би у супротном, када бисмо давали резултате за различите обиме и бројеве понављања, суштина била изгубљена и текст би постао немогућ за праћење. Ипак, напомињемо да смо, наравно, све поступке понављали за различите вредности ова два параметра, те да су резултати били скоро па потпуно исти. Такође, представићемо резултате за дводимензионалну нормалну расподелу, иако све важи и за више димензија. Слично је и са $\beta = 1$ у тест статистици. Транспарентности ради, код који смо користили за симулације јавно је доступан на налогу `github.com/cika-boske` и слободан је за употребу и модификацију.

Наша студија састоји се од два дела: у првом делу, за сваки од метода импутације пролази се кроз следећи алгоритам:

1. Генеришемо узорак из дводимензионалне[5] нормалне расподеле.

2. Извршимо генерисање MCAR недостајања у узорку, за уделе од 3% до 15%, са кораком од по пола процента, односно једног процента, у складу са рачунарским способностима.

3. За сваки од удела, у оштећеном узорку импутирамо недостајуће вредности одабраним методом.

4. Спроводимо тест, али за онај ниво значајности за који грешку прве врсте правимо у $100\alpha\%$ случајева. Такав ниво[6] зовемо *калибришуће* α.

[5]За сада, због техничких ограничења. У начелу треба испробати мноштво различитих димензија.

[6]Наравно, добије се очекивано: тестирање коришћењем оригиналног нивоа греши много више него што се очекује. Заинтересовани се охрабрују да сами покрену код.

Напомена. Из саме природе BHEP теста јасно је да параметар положаја не игра битну улогу за унимодалне расподеле. Ипак, постоји бесконачно много коваријационих матрица за које се могу генерисати подаци. Сродни су резултати за сваку од њих, а ми ћемо представити случај јединичне коваријационе матрице и матрице $\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

Алтернатива импутацији јесте брисање свих некомплетних инстанци, и са њиме желимо да поредимо наш калибрациони метод. Стога, пролазимо кроз следећи алгоритам:

1. Генеришемо узорак из неке алтернативне расподеле.

2. Извршимо генерисање MCAR недостајања у узорку, за уделе од 3% до 15%, са кораком од по пола процента, односно једног процента.

3. За сваки од удела, у оштећеном узорку импутирамо недостајуће вредности одабраним методом.

4. Такође, за сваки од удела извршимо уклањање некомплетних опсервација.

5. Поредимо емпиријске моћи за податке из два претходна корака, где за импутиране податке користимо калибришуће α.

Методи импутације које смо користили јесу импутација средњом вредношћу, SVD импутација, PMM импутација, kNN импутација и импутација коришћењем дрвета одлучивања.

Од алтернативних расподела има смисла бирати оне које је тешко детектовати, а то су оне које су на известан начин „блиске" нормалној расподели. Ми смо се одлучили за стандардне t_5, t_{10}, t_{15} расподеле, као и две мешавине нормалних расподела: прву која комбинује стандардну нормалну расподелу и нормалну расподелу са коваријационом матрицом $\mathbf{\Sigma}_1$, у односу 3 : 7 и другу која стандардну нормалну комбинује са нормалном са коваријационом матрицом $\mathbf{\Sigma}_2$ у истом односу, где је $\mathbf{\Sigma}_2 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$. Код ових мешавина, због потенцијалне бимодалности дискутоваћемо и параметре положаја. Конкретно, „нестандардном делу" алтернативе додељивали смо векторе очекивања једнаке $(0, 0)$, затим $(2, 2)$ и на крају $(5, 5)$. Ово смо радили како бисмо посматрали како се мењају моћи за различите методе при све уочљивијој бимодалности алтернативне расподеле.[7].

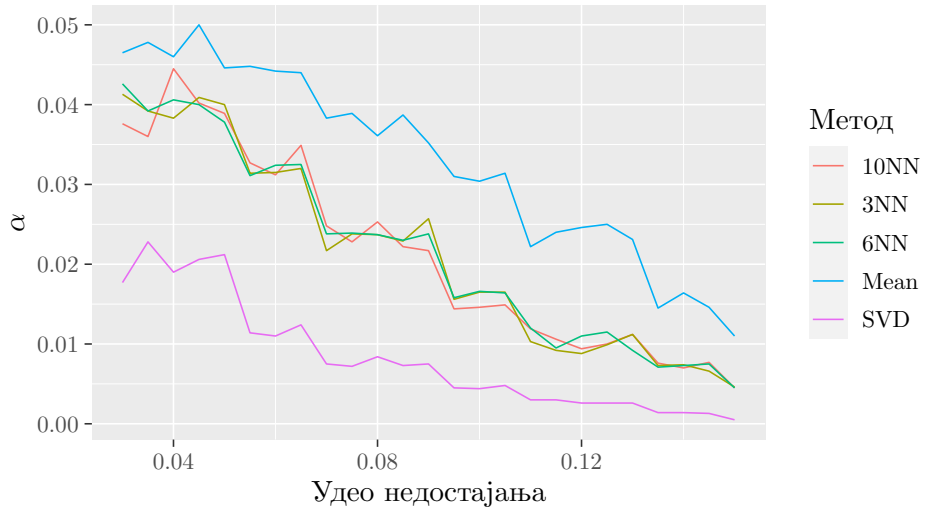
5.5.1 Резултати за узорак из стандардне нормалне расподеле

У овој ситуацији нећемо посматрати импутацију коришћењем дрвета одлучивања/случајних шума, а ево и зашто. Ми не желимо да се ограничимо на недостајања у само једној колони - то у нашем случају ни нема смисла. Када имамо недостајања у обе колоне (свим колонама), ова два метода користе MICE алгоритам, описан у примеру 4.5, који прави моделе који искоришћавају зависност међу колонама. Та зависност код нас не постоји, па нема смисла ни користити ове методе.

Стога, овде ћемо користити импутацију средњом вредношћу, SVD импутацију (дужине 1, јер то једино и има смисла пошто имамо две колоне) и импутацију алгоритмом k најближих суседа, и то за 3, 6 и 10 суседа. Номинални ниво значајности који смо користили јесте 0.05.

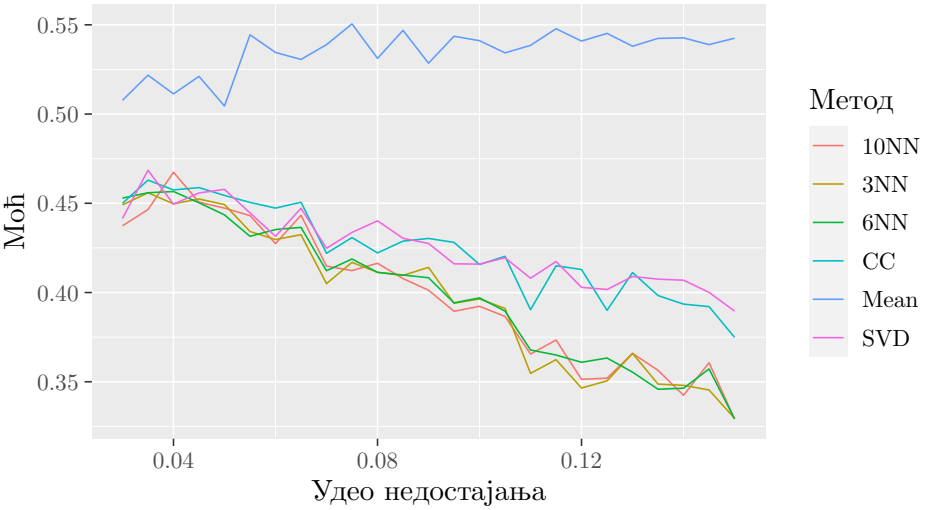
На слици 5.1 види се зависност калибришућег α од удела недостајања. Како алгоритми који су у овом случају коришћени за импутацију нису рачунски захтевни колико PMM и CART, посматрана су недостајања са кораком од по пола процента. Максималан удео недостајања који је посматран јесте 15%, јер се за веће уделе добијају нумерички нестабилне вредности за калибришући ниво значајности и оне нису употребљиве у пракси. Такође, са слике 5.1 видимо да импутација средњом вредношћу „тражи" најмање агресивну калибрацију, док се SVD импутација показала најлошије. Независно од броја суседа, сви kNN алгоритми понашају се скоро па идентично по овом критеријуму и смештени су између претходна два.

[7]Јасно, што се више разликују вектори очекивања, то је израженија бимодалност

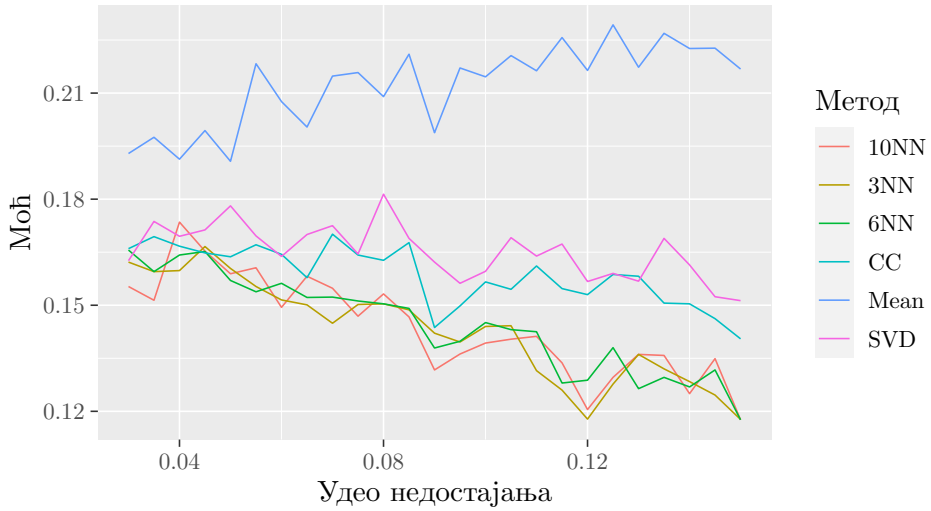


Слика 5.1: Калибришуће α за различите уделе и методе при стандардном нормалном узорку

Са слике 5.2 видимо како се при t_5, за добијене калибришуће нивое значајности, мења моћ теста у зависности од удела недостајања, при разним импутационим методима. Напоменимо да скраћеница CC одговара брисању некомплетних опсервација (енг. *complete-case*). Видимо да калибрација има најповољнији утицај на импутацију средњом вредношћу, која чак бележи и пораст моћи теста са порастом удела недостајања. Уклањање опсервација понаша се упоредиво са SVD импутацијом, док сви посматрани kNN методи бележе много већи пад моћи с порастом удела недостајања.



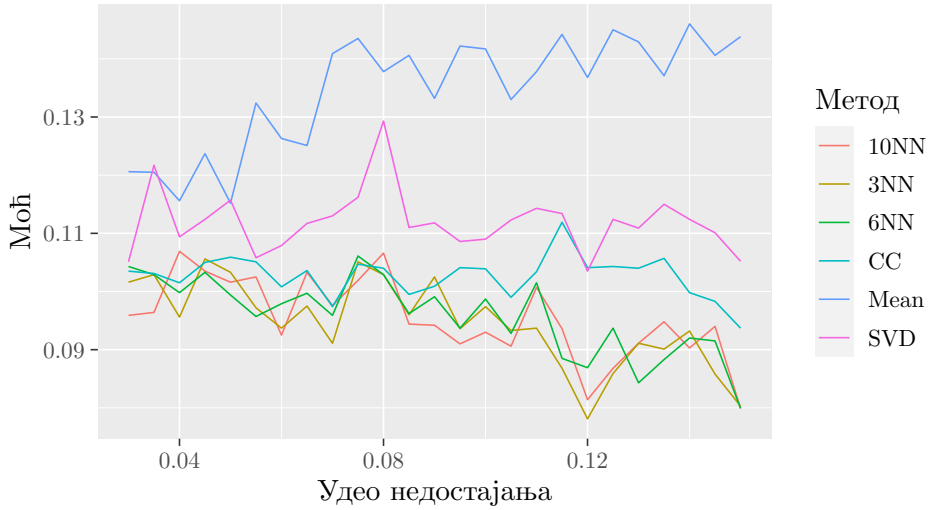
Слика 5.2: Различите моћи за t_5 алтернативу при стандардном нормалном узорку



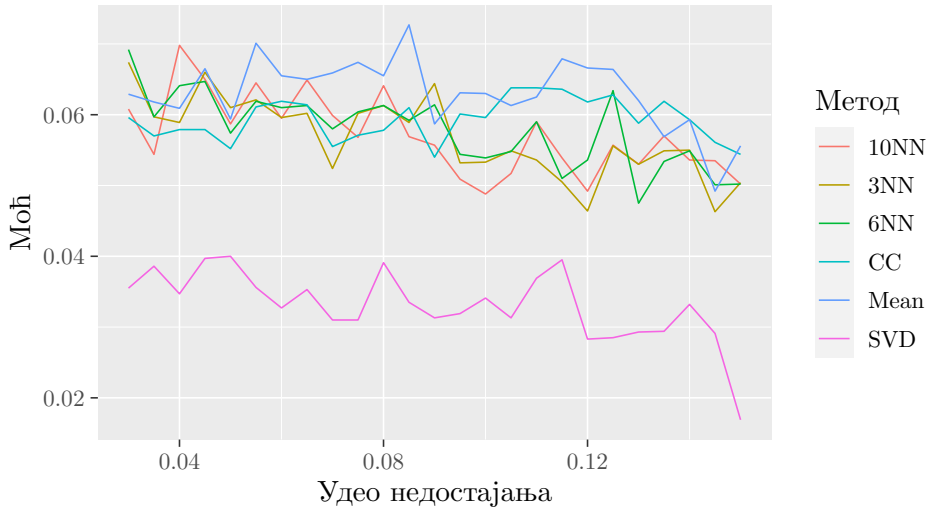
Слика 5.3: Различите моћи за t_{10} алтернативу при стандардном нормалном узорку

На слици 5.3 видимо шта се дешава са моћима у случају t_{10} алтернативе. Односи моћи за различите методе остају исти као и за t_5, али су све моћи мање. То је разумно, јер је t_{10} алтернатива по већини метрика ближа нормалној неголи t_5. Исто важи и за t_{15} алтернативу, с тим што су све моћи још мање. То се може видети на слици 5.4.

Код алтернативе која представља мешавину нормалних расподела, ситуација је другачија. За алтернативу која представља мешавину $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ на слици 5.5 већина метода даје моћ приближну номиналном нивоу значајности од 0.05, тј. сви осим SVD импутације, која даје значајно мању моћ. Ово је разумљиво сходно природи алтернативне расподеле.

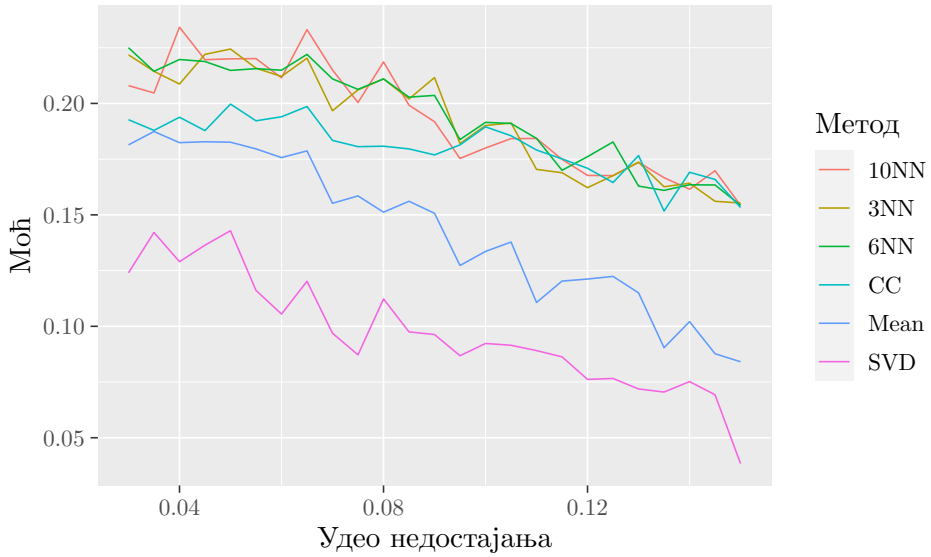


Слика 5.4: Различите моћи за t_{15} алтернативу при стандардном нормалном узорку



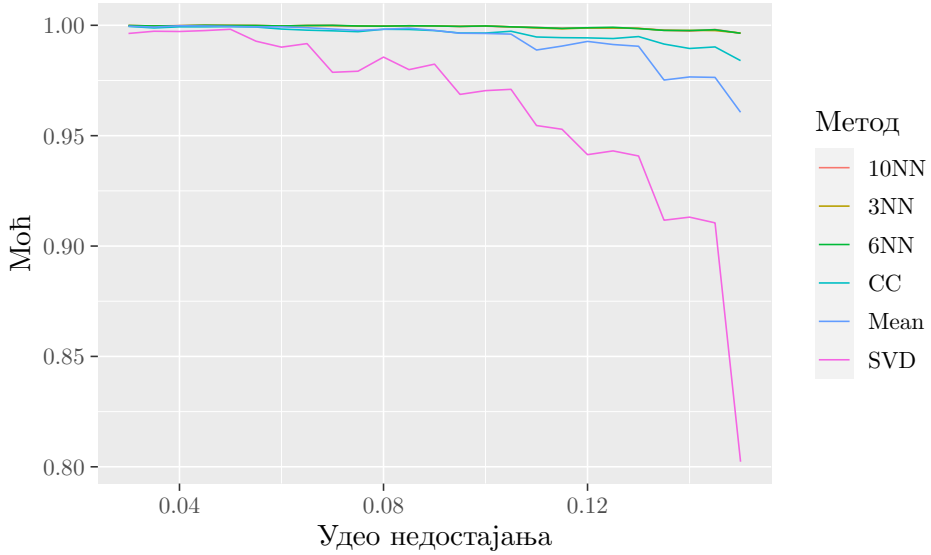
Слика 5.5: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ алтернативу при стандардном нормалном узорку

Када повећамо растојање средњих вредности расподела у мешавини и за алтернативу узмемо мешавину $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((2,2), \mathbf{\Sigma}_1)$, све моћи значајно скачу, што можемо видети на слици 5.6. SVD импутација и даље убедљиво заостаје, а занимљиво је да лоше резултате даје и импутација средњом вредношћу. То је разумљиво знајући да ће импутација средњом вредношћу тежити да наруши бимодалност. Уклањање опсервација упоредиво је са свим kNN методима, а за мање уделе недостајања значајно је лошије. Одавде можемо закључити да kNN импутација има тенденцију да очува расподелу података.

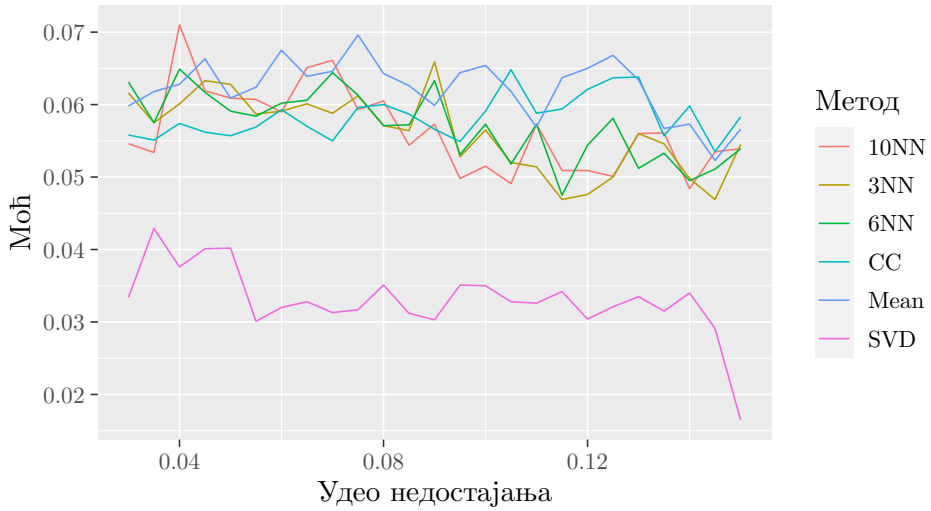


Слика 5.6: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((2,2), \mathbf{\Sigma}_1)$ алтернативу при стандардном нормалном узорку

Коначно, када за алтернативу узмемо $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((5,5), \mathbf{\Sigma}_1)$ све моћи, осим за SVD импутацију, постају блиске јединици чак и за веће уделе недостајања. SVD импутација даје значајно лошије резултате, нарочито за већа недостајања: разлике иду и до 0.2. Све ово видљиво је на слици 5.7.

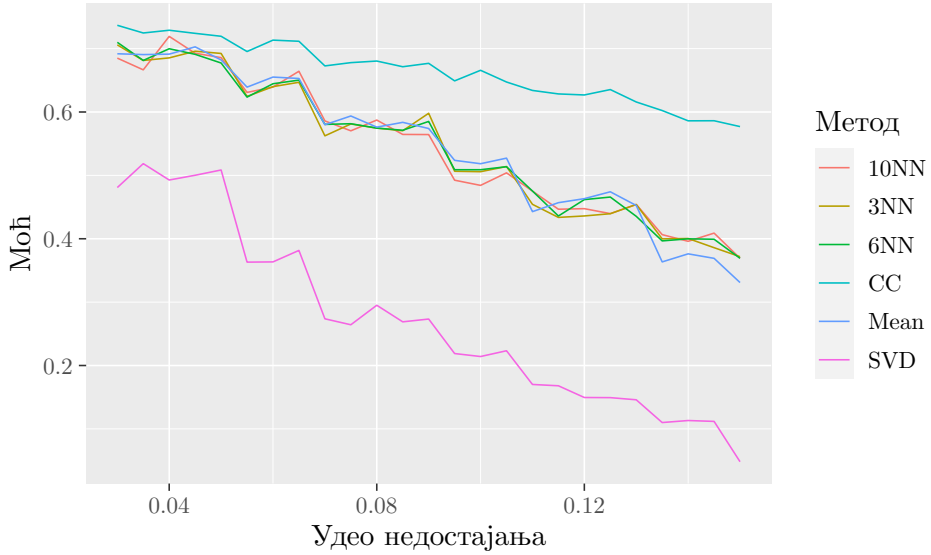


Слика 5.7: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((5,5), \mathbf{\Sigma}_1)$ алтернативу при стандардном нормалном узорку

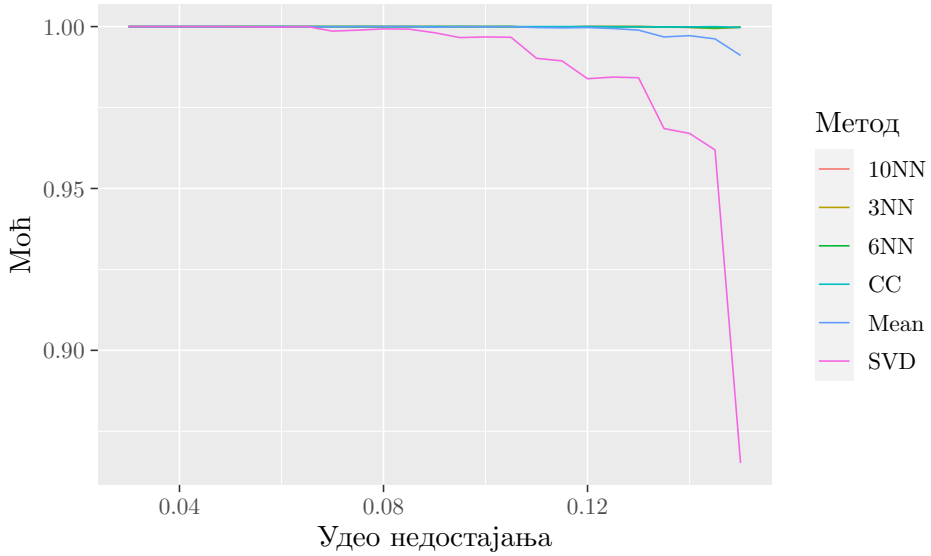


Слика 5.8: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ алтернативу при стандардном нормалном узорку

Уколико у другој расподели из мешавине матрицу $\mathbf{\Sigma}_1$ заменимо матрицом $\mathbf{\Sigma}_2$ (која, подсетимо се, на споредној дијагонали има -0.5, а не 0.5), резултати постају нешто другачији. Конкретно, SVD и даље заостаје, али уклањање опсервација добија предност у односу на импутацију средњом вредношћу и kNN. Ипак, ова предност, иако значајна, много је мања неголи у односу на SVD. То се може видети на сликама 5.8, 5.9 и 5.10, а нарочито на другој набројаној. Такође, на поменутој слици можемо видети да су све моћи значајно веће.



Слика 5.9: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((2,2), \mathbf{\Sigma}_2)$ алтернативу при стандардном нормалном узорку



Слика 5.10: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((5,5), \mathbf{\Sigma}_1)$ алтернативу при стандардном нормалном узорку

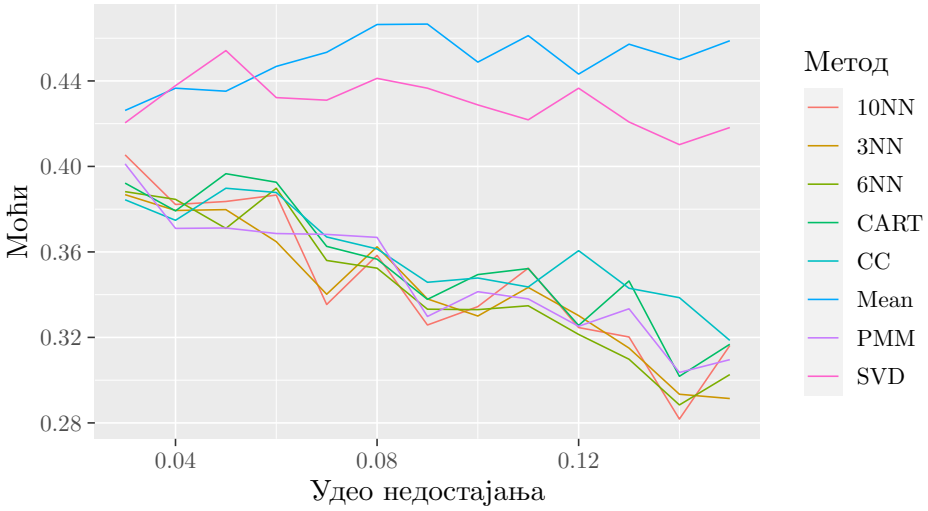
5.5.2 Резултати за узорак из нормалне расподеле са корелисаним колонама

У овом случају узорке ћемо генерисати из нормалне $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ расподеле. Како у овом случају постоји извесна корелација међу колонама, поред досадашњих метода импутације које смо посматрали, посматраћемо и PMM и CART импутацију, специјално кроз MICE алгоритам, јер не желимо да се ограничимо само на једнодимензионо недостајање. И овде ћемо се држати номиналног нивоа значајности од 0.05.

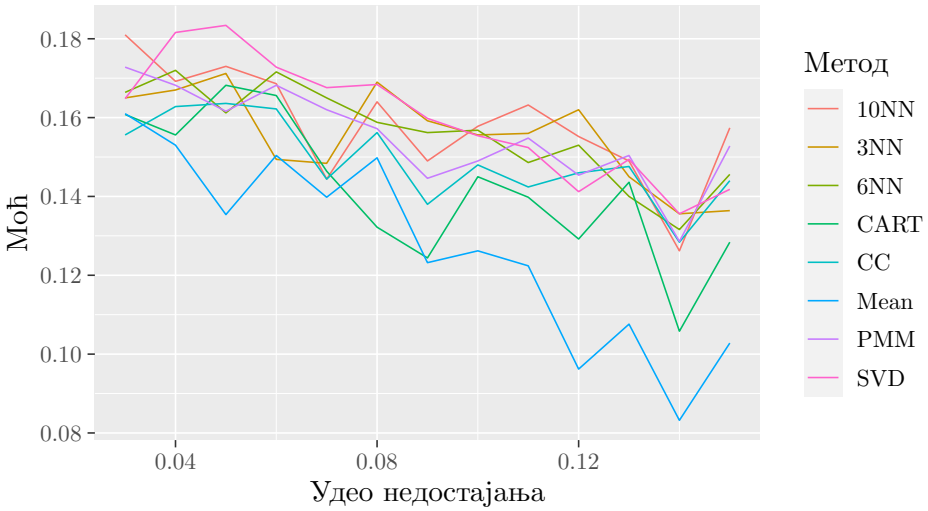
Нажалост, како се време извршавања кодова који укључују импутацију алгоритмом MICE мери данима, тренутно смо у стању да понудимо само резултате за t_5 и $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot$

$\mathcal{N}((2,2),\mathbf{\Sigma}_1)$ алтернативу. Такође, уделе недостајања и даље посматрамо у распону $3\% - 15\%$, али овај пут са кораком 1%, а не пола процента, како бисмо додатно убрзали рачунање.

Напомена. Студентова алтернатива овде је генерисана са истом коваријационом матрицом као и подаци из нулте хипотезе (матрицом $\mathbf{\Sigma}_1$), како би се тесту додатно „отежао посао" то јест да би алтернатива била што ближа нултој расподели.



Слика 5.11: Различите моћи за t_5 алтернативу при узорку из $\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_1)$ расподеме



Слика 5.12: Различите моћи за $0.3 \cdot \mathcal{N}(\mathbf{0},\mathbf{Id}) + 0.7 \cdot \mathcal{N}((2,2),\mathbf{\Sigma}_1)$ алтернативу при узорку из $\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_1)$ расподеле

За Студентову t_5 алтернативу добијају се веома занимљиви резултати.Са слике 5.11 на првом месту по успешности јесте импутација средњом вредношћу, као и код стандардног нормалног узорка, иако са нешто мањом моћи него тамо. У односу на њу, kNN методи понашају се исто као раније. Међутим, корелисаност колона довела је до тога да SVD импутација продукује значајно бољу моћ него за стандардан нормалан узорак. Изненађујуће, PMM и

CART импутација упоредиве су са kNN и уклањањем опсервација, иако би се можда од њих очекивали бољи резултати сходно њиховој природи.

За $0.3 \cdot \mathcal{N}(\mathbf{0}, \mathbf{Id}) + 0.7 \cdot \mathcal{N}((2,2), \mathbf{\Sigma}_1)$ алтернативу, генерално говорећи све моћи су незнатно мање него за узорак из стандардне нормалне расподеле, али им се однос значајно разликује, о чему сведочи слика 5.12. SVD импутација овде се пробија на сам врх по моћи, а како се слично понашала и за t_5 алтернативу можемо закључити да је овај метод у стању да добро искористи корелацију колона коју овде има на располагању. Импутација средњом вредношћу се овде, разумно, није добро показала, а остали методи упоредиви су са SVD.

5.6 Општи закључак

Наша симулациона студија је показала да је, под претпоставком непостојања везе међу колонама и претпоставком унимодалности[8], најразумније користити искалибрисану импутацију средњом вредношћу, која се показала као убедљиво најбоље решење. Следећи по квалитету јесу SVD уклањање некомплетних опсервација, који су се најлошије показали. При претпоставци независности колона и при алтернативи која је мешавина нормалних расподела, импутација средњом вредношћу није добро решење, јер јој квалитет нагло пада уколико је бимодалност алтернативе све израженија. Слично, лоше (неупотребљиво, моћ је мања од 0.05) се показала и SVD импутација. Стога, овде се препоручује kNN или, евентуално, уклањање некомплетних опсервација. Ми свакако препоручујемо импутацију, јер чува комплетне ћелије за неку даљу анализу.

Уколико се подаци генеришу из расподеле где имамо извесну корелисаност колона (изражену коваријационом матрицом $\mathbf{\Sigma}_1$), могу се посматрати и два додатна алгоритма: PMM и CART. За Студентову алтернативу са 5 степени слободе они су у истој класи по моћи као и уклањање опсервација и kNN, док се импутација средњом вредношћу поново појављује на врху, а у стопу је прати SVD импутација. Пређемо ли у овом случају на бимодалну расподелу, импутација средњом вредношћу поново постаје неупотребљива, али су овај пут сви остали методи упоредиви.

Коначно, да сумирамо: при независности колона и унимодалности препоручујемо искалибрисану импутацију средњом вредношћу (то јест искалибрисано тестирање након импутације њоме), а при одсуству унимодалности ипак је најбоље користити kNN. При постојању зависности колона, за унимодалне расподеле опет је препорука импутација средњом вредношћу, мада се може користити и SVD, ако због других ствари више одговара. За бимодалне расподеле може се користити било шта осим импутације средњом вредношћу.

Напомена. Како нам се испоставило да су PMM и CART импутација у нашем контексту калибрације нивоа значајности упоредиве са kNN методом, ми их не препоручујемо за употребу јер су рачунски неупоредиво захтевније. Примера ради, код који рачуна моћи за 3NN метод извршава се неких 20 минута, док се код који рачуна моћи за PMM извршава од неколико сати до неколико дана, у зависности од прослеђеног обима узорка, броја понављања и посматраних удела недостајања.

5.7 Даљи правци истраживања

Како је утицај импутационих метода на тестове сагласности, а нарочито нормалности, недовољно проучавана област, простор за напредак је велики. У нашем случају, први наредни корак представљао би проширивање симулационе студије, која би укључила више нултих и више алтернативних расподела. Такође, неопходно би било пронаћи начин да се посматрају недостајања већа од 15%, тј. да се реши проблем нумеричке нестабилности. Додатно, неопходно је испитати шта се дешава уколико недостајање није потпуно случајно. Све такве симулације су рачунски веома захтевне и изискују много времена, тако да је за очекивати

[8]Због ограничених ресурса ми смо за унимодалну алтернативу узимали само Студентову расподелу са различитим степенима слободе, тако да би можда било боље саветовати да се импутација средњом вредношћу користи код унимодалних алтернатива са реповима тежим од репова нормалне расподеле. То је, наравно, један од праваца даљег истраживања.

постепен прилив нових резултата. Наравно, неопходно је у студију укључити више тестова, те поредити резултате.

Затим, један од праваца истраживања јесте проверити колико су на ове начине импутирани подаци употребљиви за друге намене. Конкретно, нарушеност расподеле ми смо компензовали тестирајући „вештачки" смањеним нивоом значајности. То је често, штавише скоро увек, дало добре резултате, али само у контексту тестирања нормалности, и то једним конкретним тестом. Остаје као питање то шта бива са резултатима неке друге анализе примењене над импутираним подацима.

Употреба вишеструке импутације и комбиновање p-вредности добијених тестирањем над m допуњених скупова података и даље су недовољно разјашњене процедуре. Још увек није нађен теоријски поткрепљен начин за комбиновање p-вредности, а ни за постојећа ad hoc решења није довољно испитиван квалитет. Сходно томе, и то је један од могућих праваца.

Коначно, остаје нада и за теоријско поткрепљивање емпиријских резултата, што би био резултат који би заокружио напредак у области.

Референце

[1] Бојана Милошевић (2021). *Основи статистике.* Универзитет у Београду, Математички факултет.

[2] Predrag Janičić, Mladen Nikolić (2021). *Veštačka inteligencija.* Univerzitet u Beogradu, Matematički fakultet.

[3] Марко Обрадовић (2020). *Предавања на курсу Математичка статистика.* Универзитет у Београду, Математички факултет.

[4] Bruno Ebner, Norbert Henze (2020). *Tests for multivariate normality—a critical review with emphasis on weighted L^2-statistics.* Springer, TEST29, 845-892.

[5] Roderick J.A. Little, Donald B. Rubin (2019). *Statistical Analysis with Missing Data.* Third edition. Wiley Series in Probability and Statistics.

[6] Stef van Buuren (2018). *Flexible Imputation of Missing Data.* Second edition. Chapman & Hall / CRC.

[7] Jared S. Murray (2018). *Multiple Imputation: A Review of Practical and Theoretical Findings.* Statistical Science, 33(2), 142–159.

[8] Xianchao Xie, Xiao-Li Meng (2017). *Dissecting Multiple Imputation From a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?* Statistica Sinica, 27, 1485–1545.

[9] James R. Carpenter, Michael G. Kenward (2013). *Multiple Imputation and its Applications.* A John Wiley & Sons.

[10] Shaun Seaman, John Galati, Dan Jackson, John Carlin (2013). *What is meant by "Missing at random"?* Stat. Sci. 28 (2): 257–268.

[11] Cheng Li (2013). *Little's Test of Missing Completely at Random.* The Stata Journal. 13(4):795-809.

[12] Steven K. Thompson (2012). *Sampling.* Third edition. Wiley Series in Probability and Statistics.

[13] Владимир А. Зорич (2012). *Математический анализ. Часть 1.* МЦНМО, Москва.

[14] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). *mice: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Software, Volume 45, Issue 3.

[15] Guobing Lu, John B. Copas (2004). *Missing at random, likelihood ignorability and model completeness.* Ann. Statist. 32 754-765. MR2060176

[16] George Casella, Roger L. Berger (2001). *Statistical Inference.* Second edition. Duxbury advanced series.

[17] Stef van Buuren, Karin Groothuis-Oudshoorn (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual.* Technical Report PG/VGZ/00.038, TNO Prevention and Health, Leiden.

[18] Keith Knight (2000). *Mathematical statistics*. CHAPMAN & HALL/CRC Texts in Statistical Science Series.

[19] John Barnard, Donald B. Rubin (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4):948–955.

[20] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, David Botstein (1999). *Imputing Missing Data for Gene Expression Arrays*. Technical report, Stanford Statistics Department.

[21] Daniel F. Heitjan (1997). *Ignorability, sufficiency and ancillarity*. J. R. Stat. Soc. Ser. B Stat. Methodol. 59 375-381. MR1440587

[22] Lloyd N. Trefethen, David Bau (1997). *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia.

[23] Daniel F. Heitjan, Srabashi Basu (1996). *Distinguishing "missing at random"and "missing completely at random"*. Amer. Stat. 50 207-213. MR1422070.

[24] Donald B. Rubin (1996). *Multiple Imputation after 18+ Years*. Journal of the American Statistical Association, 91:434, 473-489.

[25] Xiao-Li Meng (1994). *Multiple-Imputation Inferences with Uncongenial Sources of Input*. Statistical Science, 9(4), 538–558.

[26] Kim-Hung Li, Xiao-Li Meng, Trivellore E. Raghunathan, Donald B. Rubin (1991). *Significance levels from repeated p-values with multiply-imputed data*. Statistica Sinica, 65-92.

[27] Roderick J.A. Little (1988). *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. Journal of the American Statistical Association, 83:404, 1198-1202

[28] Donald B. Rubin (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

[29] Donald B. Rubin, Nathaniel Schenker (1986). *Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse*. Journal of the American Statistical Association, 81(394), 366–374.

[30] William G. Cochran (1977). *Sampling Techniques*, Third edition. New York: Wiley.

[31] Donald B. Rubin (1976). *Inference and missing data*. Biometrika, 63(3):581–590.

[32] David Roxbee Cox, David Victor Hinkley (1974). *Theoretical Statistics*. Springer - Science + Business Media, B.V.

[33] S.F. Buck (1960). *A method of estimation of missing values in multivariate data suitable for use with an electronic computer*. J. R. Stat. Soc. B 22: 302–306

[34] Вебсајт *Learn e-Tutorials*, на адреси: `https://learnetutorials.com/machine-learning/classification-and-regression-trees`, приступљено 12.08.2022. у 23.17

Кратка биографија аутора

Данијел Алексић рођен је 25.09.1998. године у Зворнику, Република Српска, Босна и Херцеговина, од оца Гојка Алексића (1967-) и мајке Милофинке Алексић, рођено Капетановић (1971-2021). Нема ни брата ни сестру.

Први разред основне школе „Свети Сава" завршио је у Зелињу, Зворник. Други, трећи и четврти разред завршио је у ОШ „Вук Караџић" у Рођевићу, Зворник, а остатак основношколског образовања у подручном одељењу поменуте школе у Брањеву, Зворник. Основну школу завршио је са просечном оценом 4.93. Учествовао је, а понекад и побеђивао, на разним општинским и регионалним такмичењима из математике, енглеског језика, физике и хемије.

Средњошколско образовање стекао је у гимназији „Вук Караџић" у Лозници, природно-математички смер, коју је завршио као носилац дипломе „Вук Караџић". Учествовао је, а понекад и побеђивао, на општинским и окружним такмичењима из математике и руског језика.

Математички факултет Универзитета у Београду уписао је 2017. године на студијском програму Математика, и то модул *Статистика, актуарска и финансијска математика*, а дипломирао је 21. септембра 2021. године са просечном оценом 9.47. Исте године уписао је и мастер студије на истом студијском програму и модулу.

Говори енглески и руски језик.