Мастер рад

Бајесовски приступ моделовању структурне једначине и примене у биостатистици

Милијана Свитлица

Ментор: др Марко Обрадовић
Комисија: др Бојана Милошевић и др Милан Јовановић

Математички факултет
Универзитет у Београду
Катедра за вероватноћу и статистику

Садржај

1 Увод

У медицинским, друштвеним и психолошким истраживањима, често се сусрећемо са латентним факторима, који не могу директно да се измере. Моделовање структурне једначине нуди стабилну методологију за оцењивање сложених односа међу латентним величинама, користећи шири скуп измерених величина.

Латентни конструкти могу бити: здравствено стање, гојазност и крвни притисак. Сваки од њих може бити одређен неколицином величина које могу да се измере. Мерљиве величине - систолни и дијастолни крвни притисак заједно описују латентну величину - стање крвног притиска особе. Слично томе, однос струка и кукова, у комбинацији са индексом телесне масе описују гојазност. Замислимо затим да желимо да испитамо регресиону једначину у којој крвни притисак и гојазност предиктују здравствено стање.

Моделовање структурне једначине[1] је метод који испитује регресионе односе између латентних величина - *структурни модел*, при чему је свака латентна величина одређена неколицином (корелисаних) измерених променљивих - *модел мерења*.

Постоји више разлога зашто је регресију добро радити над латентним величинама.

1. Смањује се број независних променљивих у регресионој једначини.

2. Високо корелисане измерене променљиве групишу се у виду латентне променљиве, што ублажава проблем мултиколинеарности.

3. Интерпретација оцењених односа међу латентним величинама је смисленија него при раду са измереним величинама.

Научна истраживања често изискују развој прикладног модела за обједињено испитивање низа хипотеза о утицајима разних измерених и латентних променљивих на понашање неке циљне променљиве. Научници који практично промењују МСЈ обично имају претходно

[1]Направимо терминолошко разграничење између модела и статистичког метода који се на њега односи. Прво ће се називати *модел структурне једначине*, а друго *моделовање структурне једначине*. Оба термина ћемо замењивати скраћеницом МСЈ, а по смислу реченице ће се знати да ли је у питању метод али модел.

знање о величинама које су релевантне за конкретан модел - које измерене промењиве описују које латентне величине; које латентне величине су међусобно повезане. Оцењивање модела, провера квалитета модела и упоредна анализа конкурентних модела пружају емпиријску валидацију теоријских хипотеза које је унапред формирао истраживач. Строго логички гледано, МСЈ, као и сваки статистички модел, може се користити једино за оповргавање хипотезе, а не за доказивање њене истинитости.

> „Ако је модел сагласан са реалношћу, тада су подаци сагласни са моделом. Али ако су подаци сагласни са моделом, то не имплицира да модел одговара реалности.” - K. A. Bollen

Биостатистика је грана статистике која се бави подацима који се односе на живе организме. Студије из области биологије и медицине могу бити експерименталне или (пресечне или дугорочне) опсервацијске природе. Следећа три примера илуструју разлоге за употребу модела структурне једначине у овим студијама. Наиме, често постоји потреба за дефинисањем структурног модела који садржи латентне чиниоце.

1. Утицај медикамента на ток болести често је условљен скупом спољних латентних фактора, које је тешко контролисати дизајном експеримента. Детаљном спецификацијом структуре односа свих релевантних фактора, у фази анализе прикупљених података, расте поузданост закључка о узрочности уоченог утицаја.

2. Системи које изучава епидемиологија или екологија, описују се сложеним хијерархијским односима. Зна се да је одвојено оцењивање сваког нивоа у хијерархији недовољно квалитетан приступ.

3. Откривање веза између генетских фактора и особина организама унапређује се употребом модела који укључује латентне групације и интеракције између простих опсервација.

Упркос томе што овакви разлози често постоје, истраживачи се при планирању биостатистичке методологије ретко одлучују за приступ моделовања структурне једначине. Они поједностављују своје хипотезе, што резултује слабијим закључцима. Моделовање структурне једначине је стандардни алат у истраживањима из области психологије и

социологије. Суштински латентне величине, као што су срећа, квалитет живота или стрес, се не могу непосредно измерити, али их је могуће описати неколицином погодно изабраних питања из анкете. Софтвери за оцењивање модела структурне једначине, популарни међу истраживачима друштвених наука, користе фреквенционистичке приступе оцењивања параметара. Алтернативна, бајесовска парадигма нуди одређене предности и заслужује да достигне ширу употребу.

Основни извор за материју изложену у теоријском прегледу метода у овом раду (прва 3 поглавља), је књига [1].

1.1 Линеарни модел структурне једначине

Основна верзија МСЈ је *линеарни модел структурне једначине* (ЛМСЈ). Модел се састоји из два дела

1. *Модел мерења*

$$\mathbf{y} = \mu + \Lambda\omega + \epsilon$$

2. *Структурни модел*

$$\eta = \Pi\eta + \Gamma\xi + \delta$$

Веза између једначине мерења и структурне једначине је вектор ω, који се састоји из компоненти вектора η и ξ. Величине које учествују у једначинама су дефинисане у табелама:

Симбол	**Значење**	**Димензија**
$\mathbf{y}$	Вектор измерених променљивих	$p \times 1$
μ	Вектор слободних чланова (отсечака)	$p \times 1$
ω	Вектор латентних променљивих (фактора)	$q \times 1$
ϵ	Грешка модела мерења. случајни вектор	$p \times 1$
Λ	Матрица коефицијената (оптерећења фактора) који описују утицај измерених на латентне променљиве	$p \times q$

Симбол	**Значење**	**Димензија**
η	Вектор *зависних латентних величина*	$q_1 \times 1$
ξ	Вектор *независних латентних величина*, тј. предиктора	$q_2 \times 1$
δ	Грешка структурног модела, случајни вектор	$q_1 \times 1$
П	Матрица коефицијената који описују међусобне утицаје између циљних променљивих	$q_1 \times q_1$
Г	Матрица коефицијената који описују утицај латентних предиктора на зависне латентне променљиве	$q_1 \times q_2$

Сврставању латентних променљивих (укупно $q = q_1 + q_2$) у предикторе (укупно q_2), односно зависне променљиве (q_1), приступа се тако што се примењују теоријска знања о конкретној студији и њеним циљевима. Приметимо да структурна једначина дозвољава да неке од зависних величина утичу на друге зависне величине, онда када модел обухвата више зависних величина.

Пример 1. *Развој болести бубрега*

Развој болести бубрега (KD) је зависна латентна променљива, одређена са 2 измерене величине - однос албумина и креатинина у урину (ACR) и количина креатинина у плазми (PCr). Независне латентне променљиве које утичу на KD су крвни притисак (BP) и гојазност (OB). Њихове измерене променљиве су, као што је већ разматрано, систолни крвни притисак (SBP), дијастолни крвни притисак (DBP), однос од струка до кукова (WHR) и индекс телесне масе (BMI).

У стандардним ознакама MCJ, латентне величине су $\eta = KD$, $\xi_1 = BP$ *и* $\xi_2 = OB$, *односно* $\omega = (\eta, \xi_1, \xi_2)^T$. *Измерене величине су* $\mathbf{y} = (ACR, PCr, SBP, DBP, WHR, BMI)^T$.

Структурни модел

$$KD = \mu + \gamma_1 BP + \gamma_2 OB + \delta$$

Модел мерења

$$\begin{bmatrix} \text{PCr} \\ \text{ACR} \\ \text{SBP} \\ \text{DBP} \\ \text{BMI} \\ \text{WHR} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & \lambda_{53} \\ 0 & 0 & \lambda_{63} \end{bmatrix} \begin{bmatrix} \text{KD} \\ \text{BP} \\ \text{OB} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

1.2 Претпоставке модела ЛМСЈ

За сваку опсервацију i у узорку обима n, $1 \leq i \leq n$, препоставља се да важи:

А1: Случајни вектори ϵ_i су независни и једнако расподељени (i.i.d.), према расподели $\mathcal{N}(0, \Psi_\epsilon)$, где је Ψ_ϵ дијагонална матрица коваријације.

А2: Случајни вектори независних латентних променљивих ξ_i су независни и једнако расподељени, према расподели $\mathcal{N}(0, \Phi)$, где је Φ матрица коваријације.

А3: Случајни вектори δ_i су независни и једнако расподељени, према расподели $\mathcal{N}(0, \Psi_\delta)$, где је Ψ_δ дијагонална матрица коваријације.

А4: Вектори δ_i и ξ_i су независни. Вектори δ_i и ϵ_i су независни. Вектори ϵ_i и ω_i су независни.

Елементи матица коваријације Ψ_ϵ, Φ и Ψ_δ су непознати коефицијенти који се оцењују. Последице претпоставки А1-А4 су следеће:

1. Вектори η_i су такође независне једнакорасподељене величине са нормалном расподелом чија коваријациона матрица зависи од Π, Γ, Φ и Ψ_δ, јер се ради о линеарној комбинацији нормално расподељених независних вектора ξ_i и δ_i.

2. Слично, вектори ω_i су независни и једнакорасподељени, са нормалном расподелом.

3. Коначно, и вектори $\mathbf{y_i}$ морају бити нормално расподељени i.i.d. вектори.

Укратко, измерене случајне величине су по претпоставкама непрекидне, независне, једнакорасподеђене величине са нормалном

расподелом. Осим тога, ЛМСЈ не омогућује оцењивање нелинеарних веза. Нису укључуени квадратни чланови и интеракције међу величинама.

Овакве претпоставке базичне формулације модела су врло рестриктивне и нису испуњене у реалним ситуацијама, што се показује приликом анализе података. Постојање и употреба суптилнијих модела и статистичких метода је неопходно за доношење исправних и поузданих закључака у пракси.

1.3 Одређеност модела

Структура матрице оптерећења фактора Λ, такође се задаје у складу са сазнањима о величинама које чине студију, тако што се неки њени елементи фиксирају унапред, док се остали оцењују. Коефицијенти фиксирани на вредност 0 указују на то да истраживач претпоставља да не постоји веза између одговарајуће измерене и латентне величине. Од корака фиксирања вредности коефицијената зависи и интерпретација латентних величина. На пример, зна се да систолни крвни притисак нема директног утицаја на гојазност. Ако би одговарајући коефицијент био слободан и оцењен као не-нула вредност, латентна променљива на коју утичу однос од струка до кукова, индекс телесне масе и систолни крвни притисак не би могла да се тумачи као гојазност.

Штавише, ако би се за матрицу коефицијената Λ поставила општа форма са свим слободним коефицијентима, немогуће би било оценити такав МСЈ (осим ако погодно фиксирамо коефицијенте матрица коваријације, што је мање практичан поступак у скоро свим реалним ситуацијама).

За МСЈ се каже да је одређен уколико задата структура слободних коефицијената, у једначини мерења и структурној једначини, има јединствено решење. Питање одређености модела се обично испитује на конкретном примеру. Овај корак у развоју модела долази након што се изврши спецификација претпостављених односа између величина, а пре него што се приступи оцењивању параметара модела.

Често се примењује непреклапајућа структура матрице оптерећења, где свака врста матрице има тачно један слободан коефицијент, као на

пример

$$\begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & \lambda_{53} \\ 0 & 0 & \lambda_{63} \end{bmatrix}.$$

Она указује на то да свака измерена променљива утиче на тачно једну латентну величину. Даље, како би се обезбедила одређеност модела мерења, обично се по један коефицијент који се односи на једну латентну променљиву, поставља на вредност 1 (књига [1], 21. страна)

$$\begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{63} \end{bmatrix}.$$

Тако се уводи јединствена скала за латентну величину. Мање популаран приступ за идентификацију модела је да се фиксирају неки од коефицијената у матрици коваријација независних латентних променљивих Ф.

1.4 Нелинеарни модел структурне једначине

МСЈ се у литератури често дефинишу у форми која представља најједноставнији ЛМСЈ (без нелинеарних чланова), без међусобних утицаја између зависних латентних варијабли и без фиксних предиктора. У овом тексту, МСЈ подразумева нелинеарну формулацију модела, која се дефинише у овом поглављу, а чији је специјални случај ЛМСЈ.

Једначина мерења остаје иста као у линеарном моделу. Структурна једначина сада дозвољава нелинеарни однос независних латентних променљивих. Једначине модела, који подразумева присуство фиксних предиктора у структурном моделу, као и њихове интеракције са

латентним предикторима, су:

$$\mathbf{y} = \mu + \Lambda\omega + \epsilon$$

$$\eta = \Pi\eta + \Gamma\mathrm{G}(\xi) + \delta,$$

где је $\mathbf{G}(\xi) = (g_1(\xi), \ldots, g_t(\xi))^T$; $g_1, ..., g_t$ су дводимензионалне, ненула, познате, линеарно независне диференцијабилне функције; $t \geq q_2$. Димензија вектора Γ је сада $q_1 \times t$. Остале ознаке су исте као код ЛМСЈ.

Приметимо да зависне латентне променљиве η, не могу имати нелинеарне чланове у структурној једначини, већ само независне, ξ. Због тога се некад разматрају, упоредо, различите спецификације субмодела, који се разликују по питању сврставања латентних величина из ω у вектор η односно ξ.

Претпоставке нелинеарног МСЈ су идентичне А1-А4 из поглавља 3.3. Међутим, последице су другачије у односу на ЛМСЈ. Наиме, због присуства нелинеарних чланова који описују учешће латентних независних променљивих у структурном моделу, вектори ω_i и $\mathbf{y_i}$ нису нормално расподељени. Линеарна независност функција $g_1, ..., g_t$ доприноси обезбеђивању одређености модела.

Интеракције између латентних и измерених предиктора су честе у комплексним студијама. Такав је случај са интеракцијама ген-ген и ген-средина у анализи генома. Други пример, из студије која се бави дијабетесним обољењем бубрега (нефропатија), могу да буду интеракције између хемодинамичких путања и метаболичких путања (глукозе и липида), при активацији ренин-ангиотенцин сиситема.

Свеобухватни облик нелинеарног МСЈ, са интеракцијама између свих латентних предиктора, је модел са великим бројем непознатих параметара. Оваква спецификација модела препоручује се једино под условом да је узорак довољно велики, тако да је могуће добити прецизне оцене свих параметара. Како узорак у практичним применама често није довољан за оцењивање свих интеракција, приступа се развоју упоредних подмодела, који укључују различите интеракције. Идеалан развој ове стратегије је, да су оцене параметара који се понављају у више субмодела приближно једнаке.

Проблем величине скупа непознатих параметара и потреба за великим узорком још је израженија ако: у моделу мерења постоје категоричке променљиве; у подацима постоје недостајуће вредности;

претпостављају се хијерархијски односи међу опсервацијама; итд. Сви такви случајеви усложњавања изискују и додатне модификације дефиниције МСЈ.

1.5 Фиксни предиктори

Фиксни предиктори су измерене величине које директно предиктују зависне измерене величине у једначини мерења (ознака **c**), односно зависне латентне величине у структурној једначини (ознака **d**). Укључивањем фиксних предиктора у МСЈ добија се додатни ниво објашњења варијабилности система. Дозвољена је претпоставка да исти фиксни предиктор утиче на измерене променљиве које описују и независне и зависне латентне променљиве, истовремено. Илустрације ради, поменимо примере величина које се често узимају за фиксне предикторе у биостатистици. То су старост и пол особе. Фиксни предиктори могу да буду како непрекидне, тако и дискретне случајне величине.

ЛМСЈ са фиксним предикторима

$$\mathbf{y} = \mu + \mathrm{A}\mathbf{c} + \Lambda\omega + \epsilon$$

$$\eta = \mathrm{B}\mathbf{d} + \Pi\eta + \Gamma\xi + \delta$$

c је вектор димензије $m_1 \times 1$, а **d** вектор димензије $m_2 \times 1$. A је матрица коефицијената који описују утицај фиксних предиктора на измерене променљиве, димензије $p \times m_1$. B је матрица коефицијената који описују утицај фиксних предиктора на зависне латентне променљиве, димензије $q_1 \times m_2$.

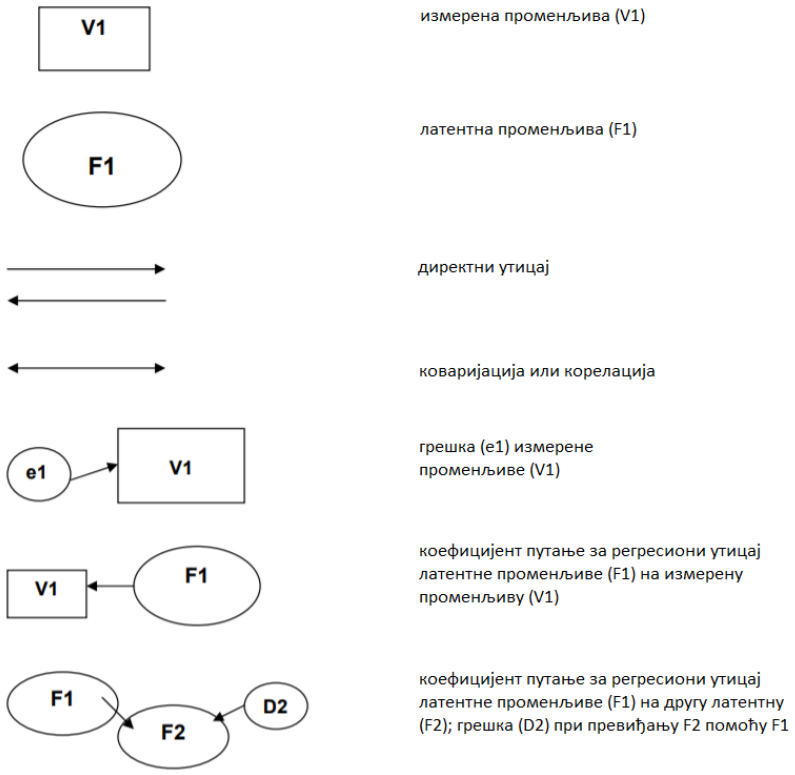
Нелинеарни МСЈ са фиксним предикторима

$$\mathbf{y} = \mu + \mathrm{A}\mathbf{c} + \Lambda\omega + \epsilon$$
$$\eta = \Pi\eta + \tilde{\Gamma}\mathrm{G}(\mathbf{d}, \xi) + \delta$$

Искоришћена је ознака $\tilde{\Gamma} = (\mathrm{B}\Gamma)$. Вектор независних променљивих G сада може да садржи и интеракције латентних величина са фиксним предикторима: $\mathbf{G}(\mathbf{d}, \xi) = (g_1(\mathbf{d}, \xi), \ldots, g_t(\mathbf{d}, \xi))^T$.

1.6 Дијаграм путање и примери из биостатистике

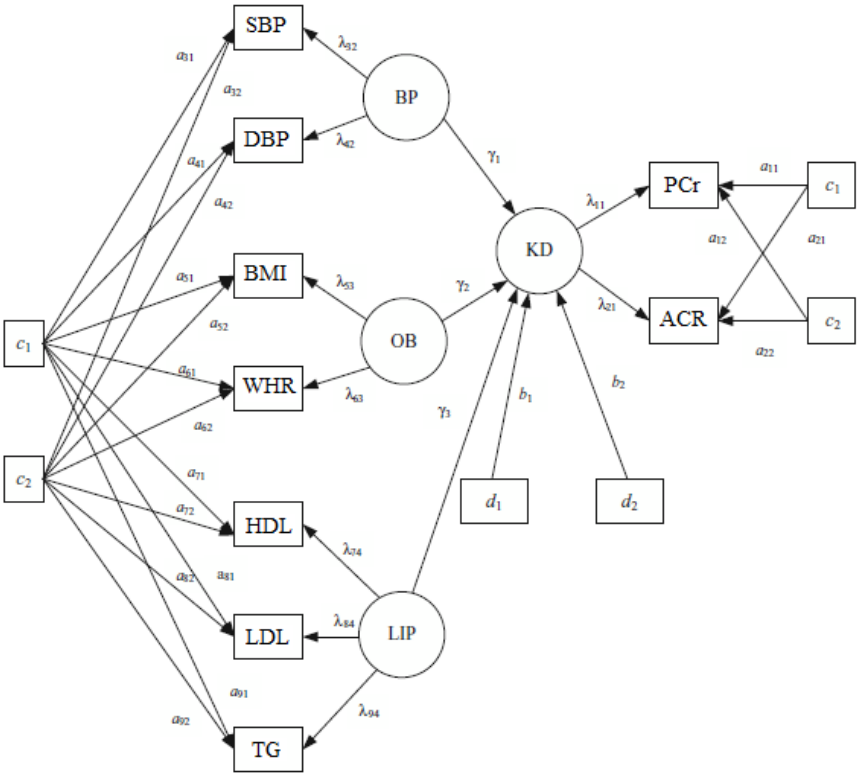
Дијаграм путање је визуелна репрезентација МСЈ - једначине мерења и структурне једначине. Ефектан је при комуникацији хипотеза и резултата везаних за студије које примењују МСЈ, али и подкласе овог модела (видети [4]).



Слика 1: Упутство за читање дијаграма путање.

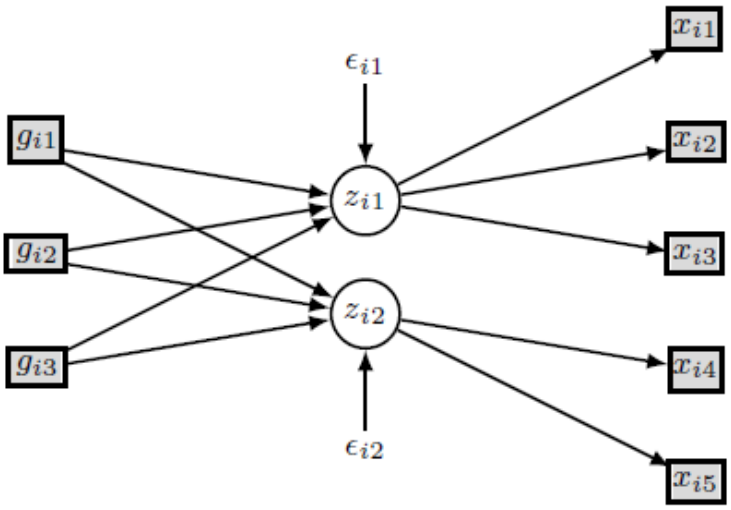
Наведимо примере конкретних МСЈ, развијених од практичара који се могу сврстати у област биостатистике.

Пример 2. *Развој болести бубрега, са фиксним предикторима*



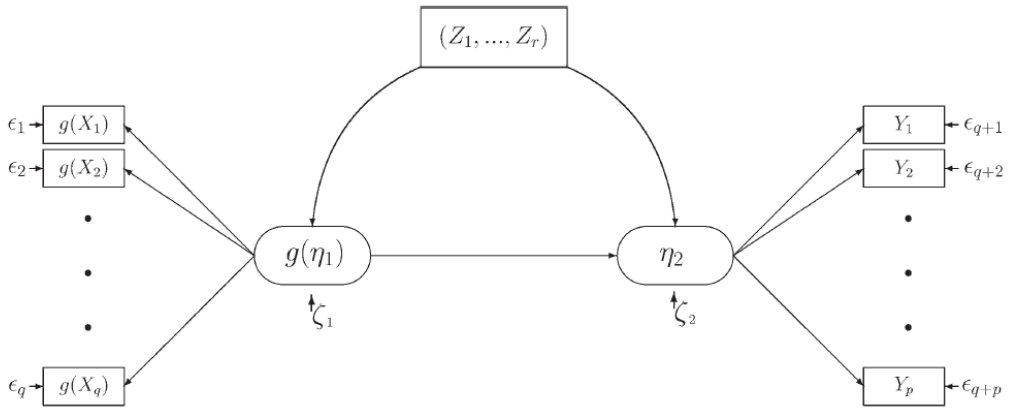
Слика 2: Дијаграм путање за MCJ који описује развој болести бубрега (KD). Независне латентне промењиве су крвни притисак (BP), гојазност (OB) и контрола липида (LIP). Величина LIP, која није већ виђена у претходном примеру, менифестује се преко 3 измерене величине - количине холестерола и триглицерида у плазми (HDL, LDL и TG). Овај модел укључује и фиксне предикторе: пушење (c_1), алкохол (c_2), старост (d_1) и пол (d_2).

Пример 3. *Статистичка анализа генома*



Слика 3: Исечак из дијаграма путање MCJ [2], који испитује генетске ризике у оквиру Иницијативе за неуроимагинацију Алцхајмерове болести. Модел користи мерења на узорку од 746 пацијената, која се састоје од 35 једнонуклеотидних полиморфизама (g_i) и 105 можданих региона (x_i). Мождани региони су разврстани у 9 групација, односно латентних величина (z_i), дефинисаних на основу знања о обрасцима груписања генетске експресије код здравих особа.

Пример 4. *Екологија*



Слика 4: Дијаграм путање MCJ из студије [3] одређивања прага безбедне изложености метил-живи из рибе у исхрани. *Функција односа доза-одговор* g, је позитивна, растућа и претпостављена унапред. Она се примењује на вектор мерења изложености метил-живи, што има ознаку X_i, за испитаника i. Сва та мерења се обједињују у виду јединствене латентне величине - изложеност $g(\eta_1)$. Изложеност директно утиче на латентну величину - одговор η_2, коју описују вишеструка мерења о последицама, Y_i. Модел такође обухвата и улогу фиксних предиктора Z_i, који имају утицаја и на зависну и на независну латентну променљиву. Показало се да методологија заснована на MCJ даје прецизнији резултат, у односу на стандардни приступ линеарне регресије.

2 Бајесов приступ оцењивању параметара МСЈ

Бајесовски приступ оцењивању коефицијената модела нуди одређене предности у односу на, популарније, фреквенционистичке начине оцењивања. Пре објашњења корака бајесовског оцењивања за МСЈ, размотримо неке паралеле између овог и фреквенционистичког приступа. Осим књиге [1], овде износимо и закључке из рада [5].

Софтверски пакети за оцењивање МСЈ се углавном базирају на принципу метода максималне веродостојности. Овај метод се ослања на претпоставке ЛМСЈ о нормалности променљивих и на достизање асимптотских својстава оцена. Да би могао да се примени за потребе МСЈ, на подацима који немају нормалну расподелу, метод је доживео разне преформулације, компликованије за разумевање и примену. Из овог разлога, МСЈ се дуго сматрао за компикован тип модела, што је значило и избегавање његове употребе.

Захваљујући достигнућима из области статистичког израчунавања, као што су ефикасни алгоритми *Марковски ланци Монте Карло* (МЛМК), започет је период шире примене бејесовског оцењивања, при анализи сложених статистичких модела. Бајесовска парадигма се не усложњава много у случају података који нису у складу са претпоставкама ЛМСЈ. Примера ради, увођењем нелинеарних веза у структурну једначину, поступак оцењивања се мења незнатно.

Фреквенционистички методи, на неки начин (нпр. метод максималне веродостојности, метод најмањих квадрата и њихове надградње), оптимизују прилагођеност оцењеног модела подацима, на глобалном нивоу. Са друге стране, Бајесов приступ подразумева генерисање опсервација из простора непознатих параметара, полазећи од доступних података и изражених предуверења о параметрима модела. Корисна импликација код МСЈ је да се генеришу и вредности латентних величина одређених моделом. Интерпретација утицаја независних латентних величина на зависне је аналогна интерпретацији регресионих коефицијената код линеарних регресионих модела. За све параметре, који се посматрају као случајне величине, из генерисаног узорка могу да се оцене како моменти тако и квантили расподеле, што је корисно при испитивању квалитета модела. Повољна је и доступност поступка анализе резидуала и детекције аутлајера.

Постоје методолошка истраживања која успешно демонстрирају идеју, да су бајесовске статистике за проверу валидности модела и поређење модела, као што је Бајесов фактор, флексибилније за коришћење и пружају одговоре на исправнија питања, у односу на класичне статистичке тестове, са p-вредностима и асимптотским расподелама тест статистика. Бајесовско тестирање хипотеза се не ослања на премису о томе да је једна од хипотеза „тачна". Аргумент против фреквенционистичког тестирања хипотеза је и феномен придавања предности алтернативној хипотези, када је узорак екстремно велик. Ипак, сличан проблем постоји и у бајесовској статистици, у виду Бартлетовог парадокса (Bartlett's paradox). Наиме, Бајесов фактор ће, неоправдано, одлучити да је модел са неинформативним априорним расподелама параметара лошији од конкурентних модела са малим дисперзијама априорних расподела. Међутим, зна се да се овај проблем превазилази тако што се, у фази поређења модела, бирају информативне расподеле, са малим дисперзијама.

Зна се да су бајесовске оцене, при великим обимима узорка, блиске оценама метода максималне веродостојности, што значи да имају иста повољна асимптотска својства. То се дешава јер, при великим узорцима, информација из узорка, садржана у функцији веродостојности, доминира над претпостављеном априорном расподелом. Међутим, истраживања су показала - случају малих узорака (релативно, у односу на број непозантих параметара), под условом да су априорне расподеле информативне, бајесовске оцене могу да буду сасвим квалитетне, за разлику од фреквенционистичких.

Најзад, у ситуацијама када истраживач поседује довољно доменског знања о феномену који моделује помоћу MCJ, пожељно је да се модел оцењује узимајући у обзир познате чињенице. Идеална ситуација би била, да је истраживачу познат резултат о односима величина које учествују у актуелној студији, из неке раније студије на сличним подацима. Управо то је главни адут бајесовског приступа. Ипак, у одсуству било каквог предуверења о односима променљивих, када се примењују неинформативне априорне расподеле, предности над класичним методима оцењивања су мање убедљиве.

2.1 Избор априорних расподела

Ефикасност алгоритама за генерисање узорка из апостериорне расподеле модела, као што је МЛМК, одакле касније следи извођење оцена параметара, зависи од типа апостериорне расподеле. Присетимо се концепта конјугованих расподела. Априорна и апостериорна расподела су конјуговане, ако припадају истој фамилији расподела вероватноћа. Знајући да априорна расподела, коју смо изабрали, има себи конјуговану, сигурни смо да апостериорна расподела има једноставну, аналитичку форму и да се из ње једноставно узоркују вредности. Стога се препоручује пракса коришћења конјугованих расподела приликом оцењивања МСЈ.

Као што је речено, бајесовска парадигма омогућује уношење сазнања, која поседује истраживач, у модел. То се имплементира у виду задавања хипер-параметара, тј. параметара априорне расподеле. На пример, за параметар положаја априорне расподеле узима се вредност коефицијента за коју истраживач претпоставља да је тачна. Затим се задаје параметар размере, којим се исказује ниво уверења о тачности одабраног параметра положаја, односно колико је дозвољено варирање око те вредности.

У одсуству претходних сазнања, могу се изводити бајесовске оцене, тако што се бирају неинформативне априорне расподеле, као што је униформна расподела. Друга опција је да се укључи информација из узорка, избором Џефрисове априорне расподеле. Уколико је доступан велики узорак, може се одвојити један део података на коме ће се, полазећи од неинформативних расподела, произвести хиперпараметри информативних априорних расподела. Затим се на преосталим подацима извршава главни поступак оцењивања МСЈ, уз примену сазнања из поменуте помоћне анализе. Такође, у случају одсуства предуверења, препоручује се спровођење анализе осетљивости модела на избор априорне расподеле и њених хипер-параметара. Дакле, стратегија за избор априорних расподела се одређује према конкретној ситуацији.

2.1.1 Стандардне конјуговане расподеле за параметре МСЈ

Непознати параметри ЛМСЈ су регресиони коефицијенти, матрице коваријација латентних величина и матрице дисперзија. Према претпоставкама модела, и измерене и латентне променљиве су

нормално расподељене. Конјугована расподела очекивања нормалне величине је нормална, док је конјугована расподела њене дисперзије инверзна гама. Ове чињенице мотивишу одлуку, да се априорне расподеле често бирају баш из фамилије нормалних и инверзних гама расподела.

Наведимо форму априорних расподела, у нелинеарном моделу са вектором отсечака и са фиксним предикторима у структурној једначини, за које се може показати да су конјуговане.

1. Модел мерења

$$\mathbf{y} = \mu + \Lambda\omega + \epsilon$$

Симбол	**Значење**	**Априорна расподела**
μ	вектор отсечака	$\mathcal{N}(\mu_0, \Sigma_0)$
$\psi_{\epsilon k}$	k-ти дијагонални елемент коварјационе матрице Ψ_ϵ	$\mathrm{IG}(\alpha_{0\epsilon k}, \beta_{0\epsilon k})$
$[\Lambda_k \mid \psi_{\epsilon k}]$	k-ти ред матрице оптерећења фактора Λ при услову $\psi_{\epsilon k}$	$\mathcal{N}(\Lambda_{0k}, \psi_{\epsilon k}\mathbf{H}_{0yk})$

$\mathrm{IG}(\cdot,\cdot)$ означава инверзну гама расподелу. [2] $\alpha_{0\epsilon k}, \beta_{0\epsilon k}$, елементи вектора μ_0 и елементи матрица Σ_0, Λ_{0k} и $\mathbf{H}_{0yk}$ су хиперпараметри. Σ_0 и $\mathbf{H}_{0yk}$ су позитивно дефинитне матрице.

2. Структурни модел

$$\eta = \mathrm{B}\mathbf{d} + \Pi\eta + \Gamma\mathbf{F}(\xi) + \delta$$

Симбол	**Значење**	**Априорна расподела**
Φ	коваријациона матрица независних латентних променљивих	$\mathrm{IW}_{q_2}(\mathbf{R}_0^{-1}, \rho_0)$
$\psi_{\delta k}$	k-ти дијагонални елемент коварјационе матрице Ψ_δ	$\mathrm{IG}(\alpha_{0\delta k}, \beta_{0\delta k})$
$[\mathrm{B}_k \mid \psi_{\delta k}]$	k-ти ред матрице B при услову $\psi_{\delta k}$	$\mathcal{N}(\mathrm{B}_{0k}, \psi_{\delta k}\mathbf{H}_{0Bk})$
$[\Pi_k \mid \psi_{\delta k}]$	k-ти ред матрице Π при услову $\psi_{\delta k}$	$\mathcal{N}(\Pi_{0k}, \psi_{\delta k}\mathbf{H}_{0\Pi k})$
$[\Gamma_k \mid \psi_{\delta k}]$	k-ти ред матрице Γ при услову $\psi_{\delta k}$	$\mathcal{N}(\Gamma_{0k}, \psi_{\delta k}\mathbf{H}_{0\Gamma k})$

[2] Случајна величина има инверзну гама расподелу са параметрима α и β, акко њена реципрочна вредност има гама расподелу, са иста та два параметра.

$\mathrm{IW}_{q_2}(\cdot,\cdot)$ означава q_2-димензионалну инверзну Вишхартову расподелу. [3] $\alpha_{0\delta k}, \beta_{0\delta k}, \rho_0$, као и елементи матрица $\mathbf{R}_0^{-1}, \mathrm{B}_{0k}, \mathbf{H}_{0Bk}, \Pi_{0k}, \mathbf{H}_{0\Pi k}, \Gamma_{0k}$ и $\mathbf{H}_{0\Gamma k}$ су хиперпараметри. $\mathbf{R}_0^{-1}, \mathbf{H}_{0Bk}, \mathbf{H}_{0\Pi k}$ и $\mathbf{H}_{0\Gamma k}$ су позитивно дефинитне матрице.

2.2 Марковски ланци Монте Карло за оцењивање модела структурне једначине

Ако искористимо устаљену нотацију, за оцену скупа непознатих параметара $\boldsymbol{\theta}$, у Бајесовој статистици, најчешће се узима математичко очекивање апостериорне расподеле $p(\boldsymbol{\theta}|\mathbf{Y})$, где је $\mathbf{Y}$ ознака за узорак. Примењује се Бајесова теорема, како би се густина апостериорне расподеле разложила на две познате компоненте - густину априорне расподеле и функцију веродостојности.

Како је поменуто очекивање, у случају МСЈ, компликован или нерешив интеграл, уместо њега се узима средња вредност узорка генерисаног из апостериорне расподеле. Дакле, тежина поступка оцењивања се преводи у проблем симулирања опсервација из апостериорне расподеле. У случају МСЈ који одступају од претпоставки најједноставнијег ЛМСЈ, апостериорна расподела се не може једноставно одредити аналитички, па су за симулирање неопходни софистицирани алгоритми.

Специфичност МСЈ, је и присуство латентних величина. Вредности латентних величина, које одговарају тачкама узорка, су такође

[3] Вишхартова расподела је вишедимензионално уопштење гама расподеле. Нека је $\mathbf{R}$ позитивно дефинитна матрица димензије $q \times q$ и нека је $\rho > q-1$. Гама функција димензије q дефинише се као

$$\Gamma_q(a) = \int_{S>0} e^{-\operatorname{tr}(S)} |S|^{a-(q+1)/2} d_S.$$

Густина Вишхартове расподеле са параметрима $\mathbf{R}$ и ρ, на носачу квадратних матрица x димензије q, је

$$f(\mathbf{x}) = \frac{|\mathbf{x}|^{(\rho-q-1)/2} e^{-\operatorname{tr}\left(\mathbf{R}^{-1}\mathbf{x}\right)/2}}{2^{\frac{\rho q}{2}} |\mathbf{R}|^{\rho/2} \Gamma_q\left(\frac{\rho}{2}\right)}.$$

Вишедимензионална случајна величина има инверзну Вишхартову расподелу са параметрима $\mathbf{R}^{-1}$ и ρ, акко њој инверзна случајна величина има Вишхартову расподелу са параметрима $\mathbf{R}$ и ρ.

непознате. Стога је апостериорна расподела у случају овог модела заправо $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$. Ради се о расподели непознате условне величине. Међутим, условна расподела $p(\boldsymbol{\Omega}|\theta, \mathbf{Y})$ обично може да се одреди на основу претпоставки модела. Исто важи и за расподелу $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$, јер се Ω сматра познатим, као услов. На тим закључцима се базирају методи МЛМК, као што је *Гибсов алгоритам* симулирања података. Гибсов алгоритам у свакој својој итерацији генерише, наизменично, опсервације из условних расподела једне по једне компоненте вектора $\boldsymbol{\theta}$ и $\boldsymbol{\Omega}$, под условом свих осталих компоненти. Тачније, у $(j+1)$-ој итерацији, користећи резултате из j-е итерације, генеришу се тачке:

$$
\begin{aligned}
&\theta_1^{(j+1)} \text{ из } p\left(\theta_1 \mid \theta_2^{(j)}, \ldots, \theta_a^{(j)}, \Omega^{(j)}, \mathbf{Y}\right)\\
&\theta_2^{(j+1)} \text{ из } p\left(\theta_2 \mid \theta_1^{(j+1)}, \ldots, \theta_a^{(j)}, \Omega^{(j)}, \mathbf{Y}\right)\\
&\vdots\\
&\theta_a^{(j+1)} \text{ из } p\left(\theta_a \mid \theta_1^{(j+1)}, \ldots, \theta_{a-1}^{(j+1)}, \Omega^{(j)}, \mathbf{Y}\right)\\
&\Omega_1^{(j+1)} \text{ из } p\left(\Omega_1 \mid \theta^{(j+1)}, \Omega_2^{(j)}, \ldots, \Omega_q^{(j)}, \mathbf{Y}\right)\\
&\Omega_2^{(j+1)} \text{ из } p\left(\Omega_2 \mid \theta^{(j+1)}, \Omega_1^{(j+1)}, \ldots, \Omega_q^{(j)}, \mathbf{Y}\right)\\
&\vdots\\
&\Omega_q^{(j+1)} \text{ из } p\left(\Omega_q \mid \theta^{(j+1)}, \Omega_1^{(j+1)}, \ldots, \Omega_{q-1}^{(j+1)}, \mathbf{Y}\right)
\end{aligned}
$$

За обичне ЛМСЈ, све ове густине су из нормалне, инверзне гама или инверзне Вишхартове расподеле. То значи да се опсервације лако и брзо генеришу. Што се тиче напреднијих варијанти МСЈ, за само генерисање опсервација из наведених условних расподела, које су компликованије, користи се *Метрополис-Хејстингс алгоритам*.

Доказано је да, под благим условима регуларности, након довољног броја *итерација загревања* J, расподела производа генерисаних опсервација $\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)}\right)$ прати жељену апостериорну расподелу $[\boldsymbol{\theta}, \boldsymbol{\Omega} \mid \mathbf{Y}]$. Алгоритам се иницира различитим вредностима, ради провере конвергенције генерисаних низова ка јединственој расподели, за задат број итерација. Такве провере се врше визуализацијом генерисаних низова свих непознатих параметара понаособ. Након што је достигнута конвергенција, прескаче се одређени број *итерација*

адаптације, након чега се поуздано сматра да опсервације прате апостериорину расподелу.

Једна од несавршености узорка из апостериорне расподеле, генерисаног на управо описан начин, је корелисаност узастопно извучених опсервација. Оцене добијене на основу узорка који није независан су, по правилу, мање прецизне. Практичан начин да се ублажи овај проблем је *прорећивање ланца*. Систематски се чувају само тачке узорка на размаку s, тј. опсервације са индексима $J + s, J + 2s, ..., J + T$. Обично мале вредности s пружају значајан допринос (нпр. $s = 10$ или $s = 20$).

Када је обезбеђен узорак из апостериорне расподеле, на њему могу да се рачунају уобичајене статистике. На тај начин се добијају оцене параметара и показатељи квалитета оцена. Нека је узорак $\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}\right) : t = 1, ..., T^*$. За једноставније МСЈ, често је довољан обим узорка $T^* = 3000$, иначе су потребни већи генерисани узорци. Бајесова оцена вишедимензионалног параметра $\boldsymbol{\theta}$ је узорачка средина гереризаних опсервација $\boldsymbol{\theta}^{(t)}$ из узорка:

$$\hat{\boldsymbol{\theta}} = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\theta}^{(t)}$$

Доказано је да је $\hat{\boldsymbol{\theta}}$ постојана оцена математичког очекивања апостериорне расподеле $p(\boldsymbol{\theta}|\mathbf{Y})$. Процена стандардне грешке добија се на основу узорачке дисперзије:

$$\widehat{\mathrm{Var}}(\boldsymbol{\theta} \mid \mathbf{Y}) = (T^* - 1)^{-1} \sum_{t=1}^{T^*} \left(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}\right) \left(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}\right)^T$$

Стандардне грешке, израчунате на овај начин, не могу да се користе за класично тестирање хипотеза о параметрима нормалне расподеле, јер се ради о различитим парадигмама закључивања. Бајесове оцене смислено се комбинују са бајесовским поступком испитивања валидности хипотеза. Понекад је од интереса да се, при испитивању прилагођености модела подацима, посматрају узорачки квантили, као оцена квантила апостериорне расподеле параметара.

За сваку од n тачака почетног узорка измерених величина из $\mathbf{Y}$, може се добити оцена одговарајућег вектора вредности латентних величина, као средња вредност одговарајућих вектора из генерисаног узорка

$$\hat{\omega}_i = T^{*-1} \sum_{t=1}^{T^*} \omega_i^{(t)},$$

где је $\omega_i^{(t)}$ је i-та колона матрице $\mathbf{\Omega}^{(t)}$. Показало се да важи да је емпиријска расподела оцена $\{\hat{\omega}_1, \ldots, \hat{\omega}_n\}$ блиска расподели стварних вектора вредности $\{\omega_{10}, \ldots, \omega_{n0}\}$.

3 Поређење и квалитет модела

Добра пракса налаже да се евалуира квалитет установљеног модела, као и да се дијагностикују недостаци, зарад потенцијалног кориговања и побољшања квалитета модела.

Захваљујући томе што се приликом оцењивања параметара MCJ генеришу и опсервације латентних величина, доступне су и оцене резидуала модела мерења и структурног модела. Нека су $\hat{\mathrm{A}}, \hat{\Lambda}, \hat{\boldsymbol{\omega}}_i, \hat{\Pi}, \hat{\eta}_i, \hat{\mathrm{B}}, \hat{\Gamma}, \hat{\boldsymbol{\xi}}_i$ ознаке за оцене вектора и матрица параметара. Вредности резидуала $\hat{\boldsymbol{\epsilon}}_i$ и $\hat{\boldsymbol{\delta}}_i$ које одговарају елементу полазног узорка измерених променљивих $\mathbf{y}_i$, са предикторима $\mathbf{c}_i$ и $\mathbf{d}_i$ су

$$\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \hat{\mathrm{A}}\mathbf{c}_i - \hat{\Lambda}\hat{\boldsymbol{\omega}}_i, \quad i = 1, \ldots, n$$

$$\hat{\boldsymbol{\delta}}_i = (\mathbf{I} - \hat{\Pi})\hat{\eta}_i - \hat{\mathrm{B}}\mathbf{d}_i - \hat{\Gamma}\hat{\boldsymbol{\xi}}_i, \quad i = 1, \ldots, n$$

Анализа резидуала обично подразумева исцртавање узорачке расподеле резидуала, као и односа са латентним и измереним променљивим. На тај начин се испитује испуњеност претпоставки модела о случајности, нормалности и хомоскедастичности (константности дисперзија), односно неформално уочавају аутлајери и обрасци у необјашњеној дисперзији.

Неки од показатеља квалитета модела користе *реплициране опсервације* $\mathbf{Y}^{rep} = (y_1^{rep}, ..., y_n^{rep})$, из расподеле узорка $p(\mathbf{Y}|\boldsymbol{\theta})$, која је одређена након што су параметри модела оцењени. Еуклидско растојање, или неко друго растојање, између генерисаног $\mathbf{Y}^{rep}$ и уоченог $\mathbf{Y}$, код коректно одабраног и оцењеног модела, је мала. Конкурентни модели могу да се пореде по принципу минимизовања тог растојања.

Најпопуларнија статистика за процену квалитета модела MCJ оцењеног на Бајесов начин, такође се заснива на процени сличности уочених и реплицираних опсервација. То је *апостериорно предиктивна p-вредност* (PPP-вредност), дефинисана изразом

$$p_B(\mathbf{Y}) = P\{D(\mathbf{Y}^{rep}|\boldsymbol{\theta}, \boldsymbol{\Omega}) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})|\mathbf{Y}, M_0\},$$

где је $D()$ *функција несагласности*, која изражава меру разилажења између две величине. То је генерализација функција које се минимизују приликом оцењивања параметара методом максималне веродостојности или методом најмањих квадрата (видети [7]). Међутим, таква функција

се у Бајесовој статистици рачуна у фази евалуације модела, након што су добијене оцене параметара. $p_B(\mathbf{Y})$ је вероватноћа горњег репа функције несагласности $D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})$, одређеног расподелом мере несагласности реплицираног узорка $D(\mathbf{Y}^{rep}|\boldsymbol{\theta}, \boldsymbol{\Omega})$. Може се сматрати да је оцењени МСЈ одговарајући ако је реализована PP p-вредност блиска вредности 0.5.

Када методологија оцењивања параметара модела прати бајесовску парадигму, тестирање хипотеза се преводи у проблем поређења конкурентних модела, који репрезентују H_0 и H_1. Сходно томе, представићемо најпознатије методе за поређење модела МСЈ.

3.1 Бајесов фактор

Бајесов фактор, као метрика за поређење модела M_1 и M_0, дефинише се изразом

$$B_{10} = \frac{p\left(\mathbf{Y} \mid M_1\right)}{p\left(\mathbf{Y} \mid M_0\right)}.$$

Из једначина

$$p\left(M_k \mid \mathbf{Y}\right) = \frac{p(\mathbf{Y}|M_k)p(M_k)}{p(\mathbf{Y}|M_1)p(M_1)+p(\mathbf{Y}|M_0)p(M_0)}, \qquad k = 0, 1$$

$$\frac{p\left(M_1 \mid \mathbf{Y}\right)}{p\left(M_0 \mid \mathbf{Y}\right)} = \frac{p\left(\mathbf{Y} \mid M_1\right) p\left(M_1\right)}{p\left(\mathbf{Y} \mid M_0\right) p\left(M_0\right)}$$

се види - у случају да се априори сматра да су оба модела једнако вероватни, Бајесов фактор је једнак изгледима у корист модела M_1, на основу апостериорних расподела. Реализације Бајесовог фактора не би требало да зависе од малих пертурбација хипер-параметара априорних расподела. Ово се испитује као део анализе сензитивности модела. Често се уместо Бајесовог фактора B_{10} рачуна његов логаритам, $2 \log B_{10}$. Усвојене су смернице поводом уверљивости закључка о избору бољег модела помоћу Бајесовог фактора, као у табели испод.

B_{10}	$2\log B_{10}$	Доказни материјал против $H_0\,(M_0)$
< 1	< 0	Не постоји, подржава се $H_0\,(M_0)$
$1-3$	0 to 2	Једва вредно спомена
$3-20$	2 to 6	Позитивно, у корист $H_1\,(M_1)$
$20-150$	6 to 10	Јако
> 150	> 10	Одлучујуће

Наведене смернице односе се на неугњеждене моделе. Иначе, оправдано је захтевати да $2\log B_{10}$ буде много веће од 6, као услов да се донесе одлука у корист модела M_1, угњежденог у M_0. Често има смисла да се при поређењу узме у обзир принцип парсимоније (Окамова оштрица), односно да се фаворизује једноставнији модел.

Вероватноће, у бројиоцу и имениоцу, у изразу B_{10}, могу се расписати у складу са једнакошћу

$$p\left(\mathbf{Y} \mid M_k\right) = \int p\left(\mathbf{Y} \mid \boldsymbol{\theta}_k, M_k\right) p\left(\boldsymbol{\theta}_k \mid M_k\right) d\boldsymbol{\theta}_k$$

Ради се о маргиналној расподели података. Овај интеграл се тешко одређује аналитички, уместо чега се користе разни поступци апроксимације. Један од актуелних приступа познат је под називом *узорковање путање*, који апроксимира $2\log B_{10}$. Притом се, ни у једном кораку, не користи априорна густина расподеле, што је повољна особина.

3.2 Информациони критеријуми

Још једна апроксимација метрике $\log B_{10}$, робусна на избор априорних расподела, је *Шварцов критеријум* (књига [1], 73. страна)

$$S^* = \log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_1, M_1\right) - \log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_0, M_0\right) - \frac{(d_1 - d_0)\log n}{2},$$

где су $\tilde{\theta}_1$ и $\tilde{\boldsymbol{\theta}}_0$ оцене методом максималне веродостојности (ММВ) скупова параметара $\boldsymbol{\theta}_1$ и $\boldsymbol{\theta}_0$, из модела M_1 односно M_0; d_1 и d_0 су димензије вектора θ_1 и θ_0; n је обим узорка. Доказано је да важи

$$\frac{S^* - \log B_{10}}{\log B_{10}} \to 0$$

Овај вид конвергенције се сматра за задовољавајући, под условом да је број степени слободе $d_1 - d_0$ релативно мали, за дату величину узорка n. Такође, у пракси се занемарује чињеница да се S^*, теоријски, дефинише користећи оцене ММВ, па се уместо њих у једначину убацују бајесовске тачкасте оцене. У случају великих узорака, ово је оправдано.

Једноставном трансформацијом, добија се *Бајесов информациони критеријум*

$$\mathrm{BIC}_{10} = -2S^* \cong -2\log B_{10} = 2\log B_{01}$$

Популарнија формулација исте идеје је BIC као показатељ квалитета једног модела M_k

$$\mathrm{BIC}_k = -2\log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_k, M_k\right) + d_k \log n,$$

с тим што важи $2\log B_{10} \cong \mathrm{BIC}_0 - \mathrm{BIC}_1$. Према упутствима из таблице са почетка овог поглавља, модел са мањом вредношћу BIC_k је бољи од два конкурентна модела.

Слично се дефинише и *Аикакеов информациони критеријум*

$$\mathrm{AIC}_k = -2\log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_k, M_k\right) + 2d_k$$

Мање вредности AIC_k значе тачнији модел. За разлику од BIC, реализација метрика AIC не зависи од обима узорка n и слабије фаворизује једноставније моделе, тј. са мањим димензијама скупа параметара d_k.

Поступци извођења једначина за два описана критеријума су различити. Ипак, све статистике које се називају „информациони критеријум", на неки начин, су мотивисане концептом минимизације неодређености модела (теорија информације). *Бајесовска девијација модела*, може да се дефинише као (видети [6])

$$D\left(\boldsymbol{\theta}\right) = -2\log p\left(\mathbf{Y} \mid \theta, M_k\right)$$

Први члан у изразима за AIC_k и BIC_k може се видети као $D\left(\tilde{\boldsymbol{\theta}}_k\right)$.

Уведимо ознаку за условно очекивање функције девијантности, што представља меру прилагођености модела подацима

$$\overline{D\left(\boldsymbol{\theta}\right)} = E_\theta\left\{D\left(\boldsymbol{\theta}\right) \mid \mathbf{Y}\right\}$$

Као генерализација статистике AIC развијен је *Информациони критеријум девијантности*

$$\mathrm{DIC}_k = \overline{D\left(\theta_k\right)} + p_D$$

Уместо простог пребројавања непознатих параметара (d_k у AIC_k), DIC_k користи *ефективни број параметара*, p_D

$$p_D = \overline{D\left(\boldsymbol{\theta}_k\right)} + 2\log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_k\right),$$

где је $\tilde{\boldsymbol{\theta}}_k$ бајесовска оцена од $\boldsymbol{\theta}_k$. Дакле,

$$\mathrm{DIC}_k = 2\overline{D\left(\theta_k\right)} + 2\log p\left(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_k\right)$$

Нека су $\left\{\boldsymbol{\theta}_k^{(j)}, j = 1, \ldots, J\right\}$ генерисане опсервације из апостериорне расподеле параметара. При израчунавању информационог критеријума девијантности, уместо $\overline{D\left(\boldsymbol{\theta}_k\right)}$ се узима оцена

$$-\frac{2}{J}\sum_{j=1}^{J}\log p\left(\mathbf{Y} \mid \boldsymbol{\theta}_k^{(j)}, M_k\right),$$

па се DIC рачуна лакше него BIC. Међутим, овде се користи претпоставка, да узорачка средина генерисаног узорка реализованих функција девијантности из апостериорне расподеле добро оцењује очекивање функције девијантности. Још једном, победнички модел међу конкурентним моделима је онај који има најмању вредност DIC. Правило из праксе могло би да гласи - уколико је разлика вредности ове статистике за два модела мања од 5, одлука о бољем моделу је недовољно поткрепљена.

4 Пример поступка моделовања података методом MCJ

Циљ овог поглавља је илустрација поступка од хипотезе до стабилног закључка поткрепљеног подацима, користећи MCJ. Притом је дозвољен и закључак - доступни подаци у комбинацији са одабраном методологијом нису прикладни за решавање проблема којим се бавимо.

Феномен који ћемо покушати да испитамо је утицај мера власти на преносивост вируса COVID-19. Разни аспекти пандемије вируса у 2020. години су изузетно популаран предмет изучавања. Како се ради о општем добру цивилизације, подаци о току развоја пандемије се ажурно систематизују и објављују на интернету, без ограничења приступа. Универзитет Џон Хопкинс предњачи у активности поводом изучавања вируса и пандемије. Њихове дневне временске серије регистрованих оболелих људи, као и умрлих, опорављених и активних случајева вируса, на нивоу земаља света, су драгоцен извор података.[4]. Други извор података који је искоришћен за оцењивање модела који описујемо у овом поглављу је *OxCGRT* (The Oxford COVID-19 Government Response Tracker [10]), односно дневна временска серија интензитета усвојених мера одговора на пандемију, за разне земље света. Оба извора су доступна за период од фебруара до августа 2020. године, за 146 земаља. Мере власти се могу сврстати у катогорије: суздржавање и затварање (школе, пословни објекти, јавна окупљања, карантин, саобраћај), економија (фиксна подршка, олакшавање задужења, фискалне мере), здравствени систем (кампање, одредбе о тестирању, праћење контакта, улагања у систем) и остало. Предложена су и сумарна обележја, која обједињују више мера из разних категорија. *Индекс одговора владе* (IR) је линеарна комбинација свих мера интензитета. *Индекс строгости* (IS) је комбинација свих мера суздржавања и затварања, као и мере интензитета јавних кампања ширења обавештености о вирусу (из категорије здравствени систем). Претпоставићемо да су IR и IS измерене промењљиве које дефинишу латентну промењљиву *одговор владе* (Gov). Премиса која оправдава потребу за коришћењем IR, поврх саме IS, је да присутност правила и

[4]Додатне агрегације овог извора могу се пронаћи на GitHub репозитотијуму *https://github.com/imdevskp/covid_19_jhu_data_web_scrap_and_cleaning*

одредби, које се не односе директно на изолацију грађана, утиче посредно на свест и савесност грађана поводом пандемије.

Да бисмо квантификовали утицај мера власти на сузбијање преносивости вируса COVID-19, посматраћемо временске интервале од 10 узастопних дана. Подаци који су доступни из наведених извора садрже времена регистрованих позитивних тестова, а не времена преношења вируса. Међутим, строгост мера власти је релевантно посматрати у тренутку преношења. Претпоставимо да се преношење вируса са особе А на особу Б десило пре тачно 10 дана. Постоји више сценарија по питању тренутака манифестације симптома код обе особе, као и тренутака када су оба случаја регистрована у званичним извештајима земље чији су становници А и Б. На основу извештаја Светске здравствене организације [11], најчешће се ради о следећим одредницама, премда постоји много различитих, ређе примећених, сценарија које оне не покривају. А је вирус пренео на Б током прва 3 дана од почетка својих симптома (симптоми су тада најјачи; преноси се капљично; могућа постсимптоматска и асимптоматска преносивост, али много ређе). А је регистрован као позитиван случај око 3 дана након првих симптома (тестирање и чекање на резултате теста). Инкубациони период, од заразе до првих симптома код особе Б је 5-6 дана (може бити и преко 14 дана, врло ретко). Затим се и Б тестира и чека на резултате око 3 дана. На основу ових сазнања, закључујемо да је позитиван тест особе А највероватније забележен пре 8 дана, или пре 7 или 9 дана, или евентуално 6 дана. Слично, особа Б је позитиван тест највероватније добила јуче, док су мање шансе да се то десило данас, пре 2 или 3 дана.

Суштински, измерене величине на којинма се заснива модел биће бројеви нових потврђених случајева вируса у некој земљи света, на различитим задршкама од референтног датума, као и индекси мера власти на задршци од 10 дана. Независне латентне величине су две: број особа које су некоме пренеле вирус пре 10 дана (T_r) и строгост мера владе пре 10 дана (Gov_{10}). Зависна латентна величина коју оне предвиђају је број особа које су вирус добиле пре 10 дана (T_d). Однос између величина може се представити мултипликативном једначином:

$$T_d = R_0 \cdot T_r \cdot exp(c \cdot Gov_{10}) \cdot G,$$

где је $c < 0$, док је G ознака за мултипликативну грешку оваквог модела. R_0 је стандардна епидемиолошка ознака (нпр. [12]) за *основни*

репродукциони број инфекције - процењени број секундарних случајева заразе које генерише једна типична оболела особа, у популацији где нико није имун на инфекцију и никакве мере сузбијања се не предузимају. [5]

Када узмемо логаритам са обе стране једнакости, добија се адитивна форма:

$$\log(T_d) = \log(R_0) + 1 \cdot \log(T_r) + c \cdot Gov_{10} + \log(G)$$

По претпоставкама за структурну једначину MCJ, $log(G)$ би морала да има нормалну расподелу са параметром положаја у нули. То значи да почетна мултипликативна грешка G има лог-нормалну расподелу. Грешка мултипликативног модела требало би да има очекивање једнако 1. Очекивање лог-нормалне расподеле са параметрима m и σ је $e^{m+\frac{1}{2}\sigma^2}$. Ако изједначимо последње две ставке, добија се $m = -\frac{1}{2}\sigma^2$. Одатле следи $\log(G) \sim \mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$. Пошто грешке у једначинама мерења MCJ по претпоставци имају очекивање 0, $-\frac{1}{2}\sigma^2$ ће бити оцењено као део одсечка, додато сабирку $\log(R_0)$. Тачније, адитивна грешка ће бити величина дефинисана као $\log(G) + \frac{1}{2}\sigma^2 \sim \mathcal{N}(0, \sigma^2)$.

Уведимо ознаке: $\mu = \log(R_0) - \frac{1}{2}\sigma^2$ и $\delta = \log(G) + \frac{1}{2}\sigma^2$. За латентне величине користићемо ознаке log_T_d и log_T_r. Добили смо финалну форму **структурне једначине** MCJ:

$$log\_T_d = \mu + 1 \cdot log\_T_r + c \cdot Gov_{10} + \delta$$

Да бисмо навели модел да генерише латентне величине које се тумаче као логаритми, и измерене величине морају да буду логаритми полазних бројева људи. Међутим, ово је повољан развој ситуације, јер логаритмоване вредности на узорку грубо прате расподелу блиску нормалној (што је претпоставка ЛМСЈ), док за сирова обележја то не важи.

[5] Служећи се епидемиолошким терминима, може се посматрати и једначина

$$T_d = R_e \cdot T_r,$$

где је R_e *ефективни репродукциони број*, илити, очекивани број секундарних случајева заразе које генерише једна типична оболела особа, у конкретним, тренутним условима хомогене популације. Идеја нашег модела је да производ $R_0 \cdot \exp(c \cdot Gov_{10})$ даје глобалну апроксимацију вредности R_e посматране 10 дана од референтног датума, а грешка G сакупља интерпопулацијску варијацију (између земаља света).

Нека је N_i број нових позитивних тестова који су регистровани i дана пре референтног датума. Смислена, мултипликативна веза између ових измерених променљивих и латентне T_r, може се задати на следећи начин:

$$N_i = c_i \cdot T_d \cdot G_i, \qquad i \in \{6, 7, 8, 9\}$$

где за коефицијенте мора да важи $0 < c_9, c_8, c_7, c_6 < 1$ и $c_9 + c_8 + c_7 + c_6 \leq 1$, а априори се претпоставља и однос $c_8 > c_9, c_7 > c_6$. Примера ради, интерпретација коефицијента c_9 је онда - пропорција особа које преносе заразу на задршци 10 (улога особе А у тексту изнад), чији су позитивни тестови забележени на задршци 9. Са друге стране, на сличан начин се може разложити и број особа које су заражене на задршци 10 (улога особе Б):

$$N_i = c_i \cdot T_d \cdot G_i, \qquad i \in \{0, 1, 2, 3\}$$

где за коефицијенте мора да важи $0 < c_0, c_1, c_2, c_3 < 1$ и $c_0 + c_1 + c_2 + c_3 \leq 1$, а априори се претпоставља и однос $c_1 > c_0, c_2 > c_3$. Ако применимо логартмовање на сваку појединачну једначину, мултипликативни однос прелази у адитивни:

$$log(N_i) = log(c_i) + 1 \cdot \log(T_r) + \log(G_i), \quad i \in \{0, 1, 2, 3\} \cup \{6, 7, 8, 9\}$$

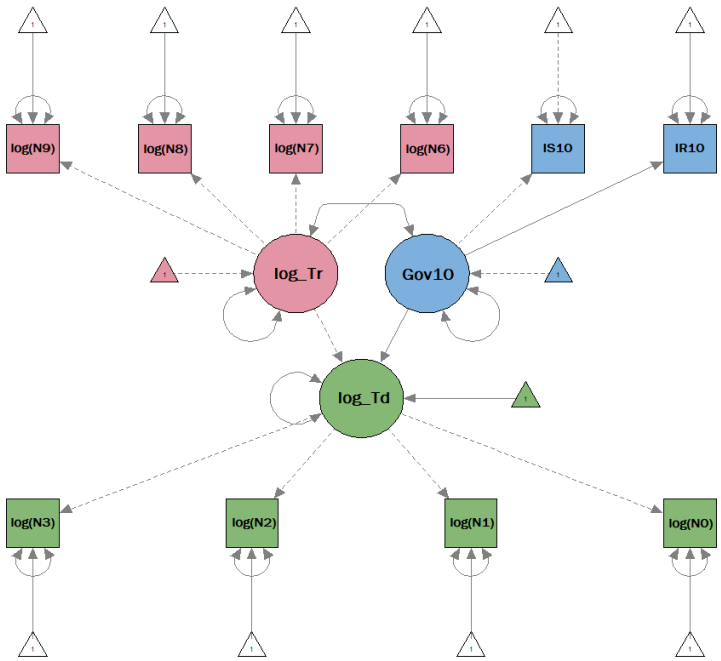
Поново желимо да мултипликативне грешке са лог-нормалном расподелом имају очекивање 1, а нормално расподељене адитивне грешке у моделу мерења MCJ имају очекивање 0. Пребацујемо очекивања адитивних грешака у одсечке једначина. За $i \in \{0, 1, 2, 3\} \cup \{6, 7, 8, 9\}$, добили смо систем

$$log(N_i) = (log(c_i) - \frac{1}{2}\sigma_i^2) + 1 \cdot \log(T_r) + (\log(G_i) + \frac{1}{2}\sigma_i^2)$$

Ради поједностављења нотације, уведимо смене $\mu_i = \log(c_i) - \sigma_i^2/2$ и $\epsilon_i = \log(G_i) + \sigma_i^2/2$. Финализовали смо поставку **једначине мерења**:

$$\begin{bmatrix} log(N_9) \\ log(N_8) \\ log(N_7) \\ log(N_6) \\ log(N_0) \\ log(N_1) \\ log(N_2) \\ log(N_3) \\ IS_{10} \\ IR_{10} \end{bmatrix} = \begin{bmatrix} \mu_9 \\ \mu_8 \\ \mu_7 \\ \mu_6 \\ \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ 0 \\ \mu_{IR} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{IR} \end{bmatrix} \begin{bmatrix} log\_T_r \\ log\_T_d \\ Gov_{10} \end{bmatrix} + \begin{bmatrix} \epsilon_9 \\ \epsilon_8 \\ \epsilon_7 \\ \epsilon_6 \\ \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_{IS} \\ \epsilon_{IR} \end{bmatrix}$$

Спецификација модела може се представити дијаграмом путање.



4.1 Априорне расподеле

Као што је објашњено, пожељно је поседовање што више предзнања о односима променљивих у моделу, што значи способност постављања информативних априорних расподела. На основу задате спецификације, која форсира конвертибилност у мултипликативне једначине, неки параметри МСЈ су фиксирани као вредност 1 или 0. За слободне параметре се могу одредити смислени домени вредности, и то са различитим претпостављеним вероватноћама. Сви слободни параметри, са одговарајућим одабраним априорним расподелама наведени су у табели.[6] Затим следе образложења резоновања о избору хипер-параметара.

Параметар	**Априорна расподела**	**Информативност**
c	$\mathcal{N}(-0.015, 0.005)$	Информативна
σ	$\mathrm{IG}(9, 4)$	Информативна
μ	$\mathcal{N}(1.125, 0.25)$	Информативна
$\sigma_0, \sigma_1, \sigma_2, \sigma_3$	$\mathbf{IG}(11, 3)$	Информативна
$\sigma_6, \sigma_7, \sigma_8, \sigma_9$	$\mathbf{IG}(11, 3)$	Информативна
μ_8	$\mathcal{N}(-0.95, 0.05)$	Информативна
μ_1	$\mathcal{N}(-0.95, 0.05)$	Информативна
μ_9	$\mathcal{N}(-1.65, 0.05)$	Информативна
μ_7	$\mathcal{N}(-1.65, 0.05)$	Информативна
μ_2	$\mathcal{N}(-1.65, 0.05)$	Информативна
μ_0	$\mathcal{N}(-1.65, 0.05)$	Информативна
μ_6	$\mathcal{N}(-2.35, 0.05)$	Информативна
μ_3	$\mathcal{N}(-2.35, 0.05)$	Информативна
μ_{IR}	$\mathcal{N}(0, 32)$	Неинформативна
λ_{IR}	$\mathcal{N}(0.7, 0.1)$	Слабо информативна
$\sqrt{\mathbf{D}(Gov_{10})}$	$\mathrm{IG}(6, 10)$	Слабо информативна
$\sqrt{\mathbf{D}(IS_{10})}$	$\mathrm{IG}(6, 10)$	Слабо информативна
$\sqrt{\mathbf{D}(IR_{10})}$	$\mathrm{IG}(6, 10)$	Слабо информативна
$\sqrt{\mathbf{D}(log\_T_r)}$	$\mathrm{IG}(6, 10)$	Слабо информативна
$cov(log\_T_r, Gov_{10})$	$\beta(1, 1)$	Неинформативна

[6] $\mathbf{D}(\cdot)$ представља оператор дисперзије. Његов корен једнак је стандардној девијацији.

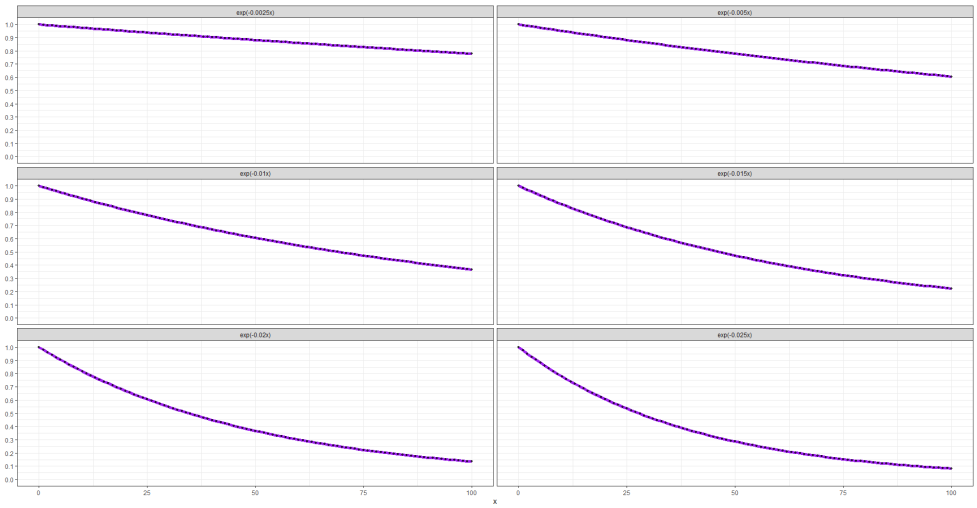
Модел ће се оцењивати као ЛМСЈ. Користићемо конјуговане типове априорних расподела - нормалне и инверзне гама.[7] Коваријација независних латентних величина имаће априорну бета расподелу.

4.1.1 Образложења

Посматрајмо део модела мерења који се односи на одредбе власти

$$\begin{bmatrix} IS_{10} \\ IR_{10} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu_{IR} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & \lambda_{IR} \end{bmatrix} \begin{bmatrix} log\_T_r \\ log\_T_d \\ Gov_{10} \end{bmatrix} + \begin{bmatrix} \epsilon_{IS} \\ \epsilon_{IR} \end{bmatrix}$$

Како смо фиксирали први одсечак на вредност 0, латентна величина Gov_{10} ће се поклапати са IS_{10}, до на одређени шум. Како се вредности IS_{10} крећу на скали $1 - 100$, и Gov_{10} ће имати сличан скуп вредности. Погледајмо сада претпостављени график функције која одређује утицај јачине мера власти на преносивост вируса.

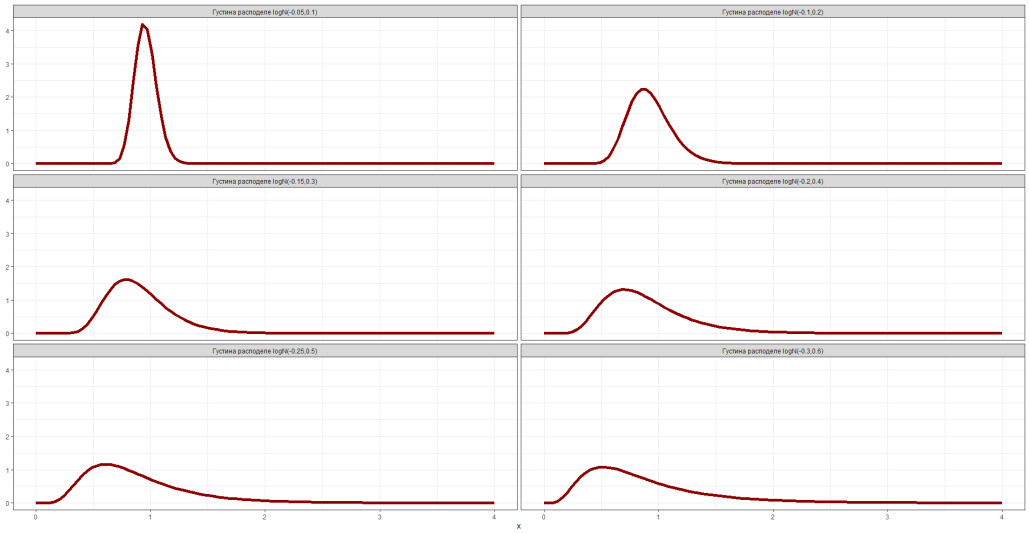


Слика 5: График функције $\exp(c \cdot x)$ на домену $[0, 100]$, за $c \in \{-0.0025, -0.005, -0.01, -0.015, -0.02, -0.025\}$

[7] R библиотеке не подржавају директно дефинисање жељене априорне инверзне гама расподеле. Уместо тога, могуће је одабрати одговарајућу гама расподелу реципрочне вредности стандардне девијације. То је еквивалентно постављању инверзне гама на параметар стандардне девијације, уместо на параметар дисперзије.

Параметар λ_{IR}, као манифестација латентне величине Gov_{10}, се вероватно налази у околини вредности 0.7. Овај параметар говори о релативној важности промењиве IR_{10} у поређењу са фиксираним коефицијентом оптерећења 1, промењиве IS_{10}. Што се тиче параметра μ_{IR}, дозвољавамо широк скуп позитивних и негативних вредности, односно бирамо неинформативну априорну расподелу. Као што видимо са графика експоненцијалних функција изнад, коефицијент c вероватно има негативну вредност, већу од -0.025. У супротном би интерпретација била - при најоштријим мерама власти, преносивост вируса је готово 0, што није реалистично.

Зна се да је дисперзија лог-нормално расподељене случајне величине једнака $e^{2(m+\sigma^2)} - e^{2m+\sigma^2}$. Штавише, на основу визуелизације густина лог-нормалних расподела $\log \mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$, варирајући параметар σ^2, бирамо хиперпараметре априорне расподеле параметра σ^2.



Слика 6: Густине расподеле $\log \mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$ расподеле, за вредности $\sigma^2 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$

Притом имамо у виду улогу мултипликативне грешке G са том лог-нормалном расподелом, али и улогу параметра σ^2 као дисперзије зависне промењиве структурне једначине. Овај избор даље утиче на избор априорне расподеле одсечка $\mu = \log(R_0) - \frac{1}{2}\sigma^2$, који је такође важан у смислу интерпретабилности решења. Параметар дисперзије,

односно стандардне девијације, (структурне једначине) се обично бира тако да одражава ниво уверења поводом избора осталих априорних расподела у једначини. При високом уверењу о адекватности информативних расподела, може да се користи инверзна гама расподела са очекивањем 0.5, што је нпр. $\mathrm{IG}(9, 4)$ (препоручени хиперпараметри у књизи [1]), као априорна расподела стандардне девијације. Ту вредност комбинујемо са R_0 приликом избора априорне расподеле одсечка. Епидемиолошка литература (нпр. [9]) наговештава да је валидан опсег вредности основног репродукционог броја R_0 за вирус COVID-19 између 2 и 8. Највећу тежину би требало да има $R_0 \approx 3$ (нпр. [9]), што повлачи $\log(R_0) \approx 1$, а онда и $\mu \approx 1.125$.

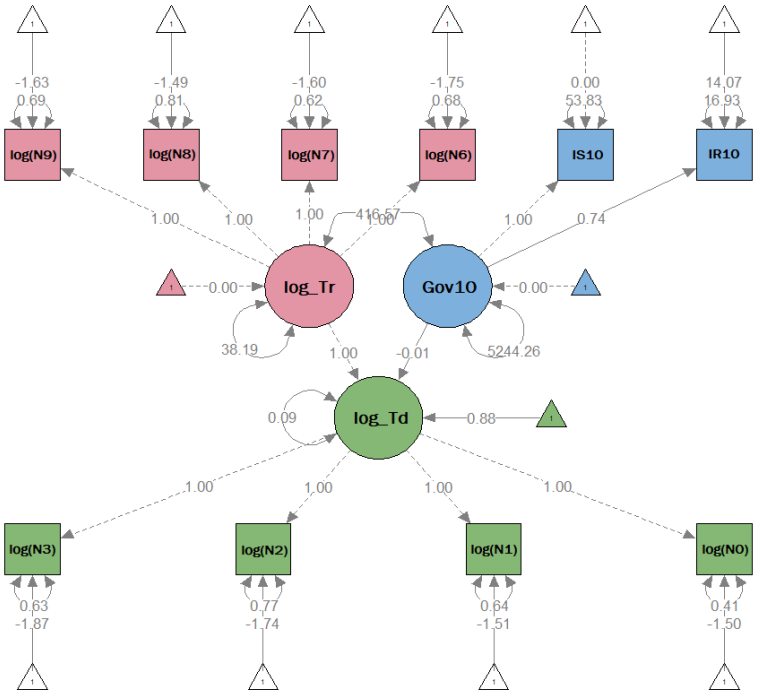
Претпоставићемо да су логаритми бројева регистрованих случајева вируса прецизно одређени једначином мерења. Ако за стандардне девијације измерених променљивих изаберемо априорну расподелу $\mathrm{IG}(11, 3)$, очекивања дисперзија су 0.09. То значи да су полазне мултипликативне грешке густо сконцентрисане око вредности 1 (видети горе наведене графике лог-нормалних расподела). Претпоставимо односе коефицијената на следећи начин: $c_1 = c_8 = 0.4$, $c_0 = c_2 = c_9 = c_7 = 0.2$ и $c_3 = c_6 = 0.1$, у складу са њиховом улогом при интерпретацији (када је регистрован позитиван тест особа А и Б). Априори претпостављене вредности одговарајућих одсечака су онда $\mu_i \approx \log(c_i) - \frac{0.09}{2}$, односно нека од вредности $-0.95, -1.65$ и -2.35.

4.2 Оцењени модел

Опсервације се односе на дан у години у једној земљи света. Опсервације за једну земљу су корелисане. Да би доступни подаци били прикладни за било какву статистичку анализу, потребно је да се апроксимира независност опсервација. Зато је модел оцењиван на случајном узорку обима 1000, извученом из целокупног скупа података. У одељку о анализи сензитивности биће испитан утицај конкретног узорка на забележени резулат.

Сва изачунавања извршена су унутар софтверског окружења R, првенствено помоћу библиотеке *blavaan* (видети [8]). Имплицитно се позива пакет *rstan*, у фази извршавања МЛМК симулација. Stan је пројекат који имплементира алгоритам *„No-U-Turn" Хамилтонови*

МЛМК. Ради се о савременом, комплексном алгоритму (видети [13]). Појекат се води императивом компјутерске ефикасности, па је главни код писан у језику C++. Покушај примене алгоритма који је верзија Гибсовог метода, описаног у овом раду, позивањем библиотеке *rjags*, показао се као лошија стратегија. За исцртавање дијаграма путање коришћен је пакет *semPlot*.



Слика 7: Дијаграм путање оцењеног модела.

Покренуте су МЛМК симулације за 3 ланца, што ће пружити увид у квалитет конвергенције ка апостериорним расподелама параметара модела. Ланци пролазе кроз 8 хиљада итерација загревања, затим се

прескаче 2 хиљаде итерација адаптације, коначно, узима се 8 хиљада опсервација за које се претпоставља да припадају апостериорним раподелама.

Оцене параметара су представљене дијаграмом путање и следећом табелом.

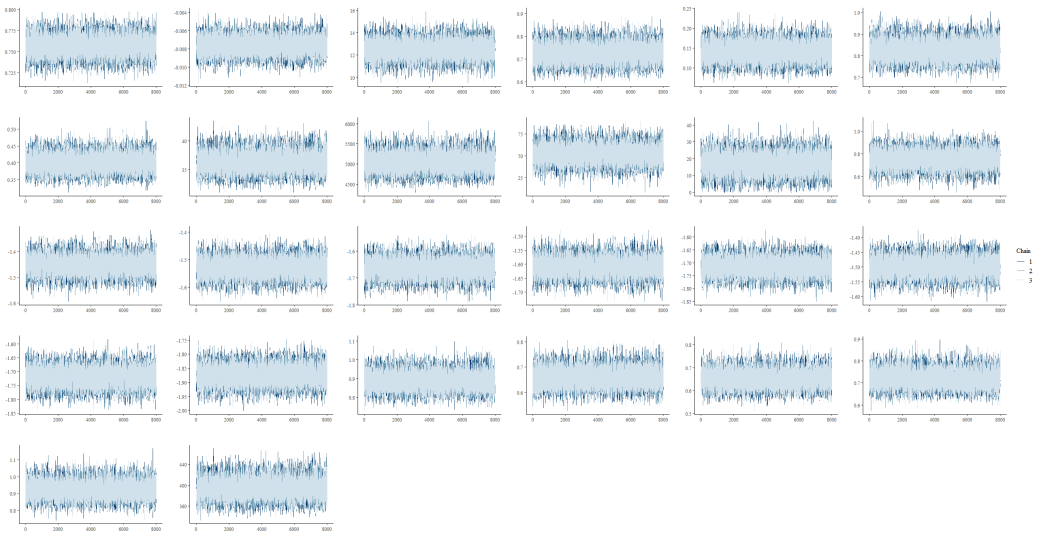
Параметар	**Оцена**	**Стандардна девијација**	**Инntервал покривања**
c	-0.009	0.001	[-0.011,-0.007]
σ^2	0.153	0.017	[0.122,0.187]
μ	0.883	0.079	[0.728,1.038]
σ_0^2	0.403	0.024	[0.358,0.451]
σ_1^2	0.625	0.034	[0.562,0.695]
σ_2^2	0.754	0.039	[0.681,0.833]
σ_3^2	0.616	0.033	[0.554,0.684]
σ_6^2	0.663	0.037	[0.593,0.739]
σ_7^2	0.606	0.033	[0.546,0.673]
σ_8^2	0.782	0.042	[0.705,0.867]
σ_9^2	0.671	0.035	[0.605,0.743]
μ_8	-1.493	0.033	[-1.558,-1.429]
μ_1	-1.509	0.032	[-1.571,-1.446]
μ_9	-1.633	0.032	[-1.696,-1.571]
μ_7	-1.599	0.032	[-1.661,-1.536]
μ_2	-1.740	0.032	[-1.802,-1.679]
μ_0	-1.495	0.030	[-1.554,-1.435]
μ_6	-1.746	0.032	[-1.809,-1.683]
μ_3	-1.865	0.032	[-1.928,-1.802]
μ_{IR}	13.694	0.841	[12.047,15.349]
λ_{IR}	0.743	0.012	[0.719,0.767]
$\mathbf{D}(Gov_{10})$	5341.767	240.494	[4886.48,5837.047]
$\mathbf{D}(IS_{10})$	61.344	11.591	[38.009,80.904]
$\mathbf{D}(IR_{10})$	11.956	6.481	[1.97,25.101]
$\mathbf{D}(log\_T_r)$	38.761	1.752	[35.473,42.335]
$cov(log\_T_r, Gov_{10})$	423.853	19.740	[386.429,464.29]

4.3 Контролна листа бајесовске дијагностике

Позивамо се на рад [14], у коме се дефинишу кораци за извештавање резултата у бајесовској статистици, у циљу достизања неопходног стандарда методолошке транспарентности. Прва тачка, *Разумевање априорних расподела*, обрађена је у поглављу 4.1.

2. *Ковергенција МЛМК ланаца.*

Ова тачка поткрепљена је графицима испод.



Слика 8: График ланаца МЛМК кроз итерације (traceplot) показује да је наступила конвергенција ка истој расподели за сва 3 ланца, за све параметре модела.

Такође, уграђене метрике пакета blavaan, *Rhat* (пореди оцене добијене у 3 одвојена ланца) и $neff$ (ефективни обим узорка, говори о утицају корелисаности генерисаних апостериорних опсервација), обе потврђују да је дошло до конвергенције МЛМК алгоритма.

3.*Могућност случаја локалне конвергенције*

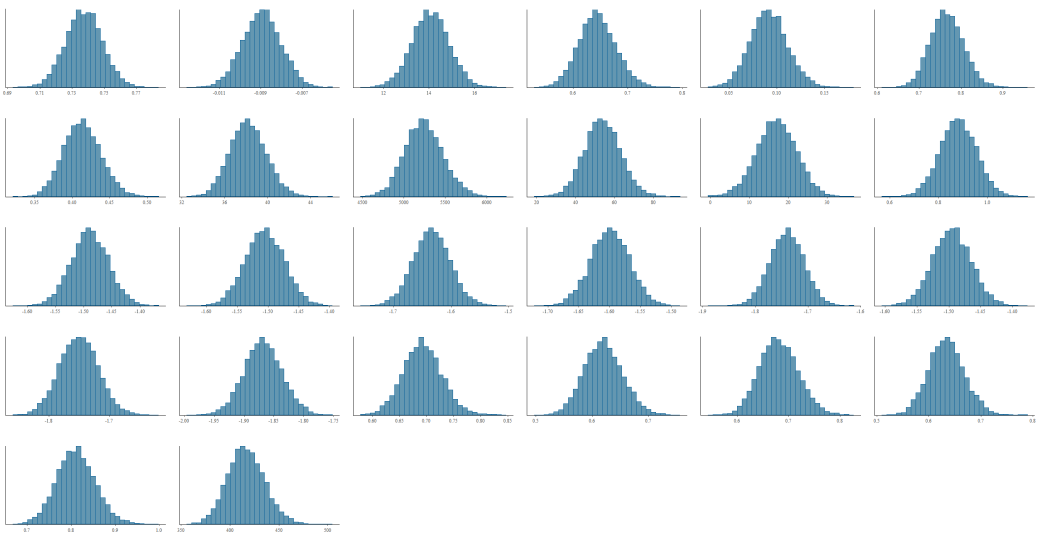
Оценићемо исти модел користећи дупло већи број итерација загревања - 16 хиљада. Оцене параметара (колона „Оцене burn-in 16k" у табели испод) поредимо са иницијалним оценама, из поглавља 4.2. Изостављамо оцене које су непромењене, на трећој децималној цифри.

Параметар	**Оцена**	**Оцена burn-in 16k**
σ_2^2	0.754	0.753
σ_8^2	0.782	0.783
μ_2	-1.740	-1.741
μ_{IR}	13.694	13.682
λ_{IR}	0.743	0.744
$\mathbf{D}(Gov_{10})$	5341.767	5340.038
$\mathbf{D}(IS_{10})$	61.344	61.498
$\mathbf{D}(IR_{10})$	11.956	11.849
$\mathbf{D}(log\_T_r)$	38.761	38.757
$cov(log\_T_r, Gov_{10})$	423.853	423.730

Видимо да се ни једна од оцена параметара није значајно променила. Не ради се о локалној конвергенцији, што смо и хтели да покажемо.

4. *Хистограми апостериорних расподела.*

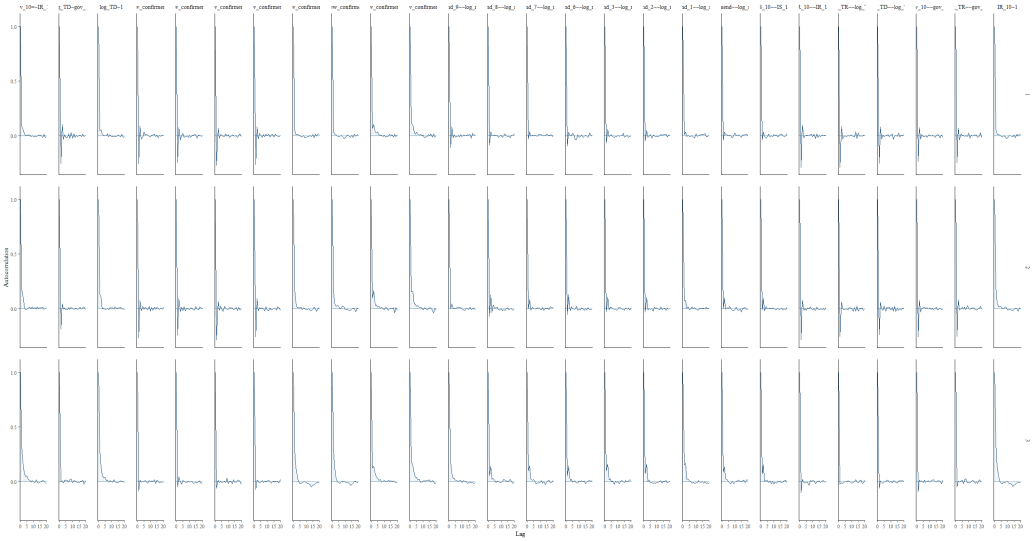
Да ли је генерисани обим узорка довољно велик да реконструише информацију о целој апостериорној расподели? Да ли постоје неправилности и празнине на хистограмима?



Слика 9: Сви хистограми апостериорних расподела параметара су правилни, са довољно детаља.

5. *Аутокорелације унутар ланаца.*

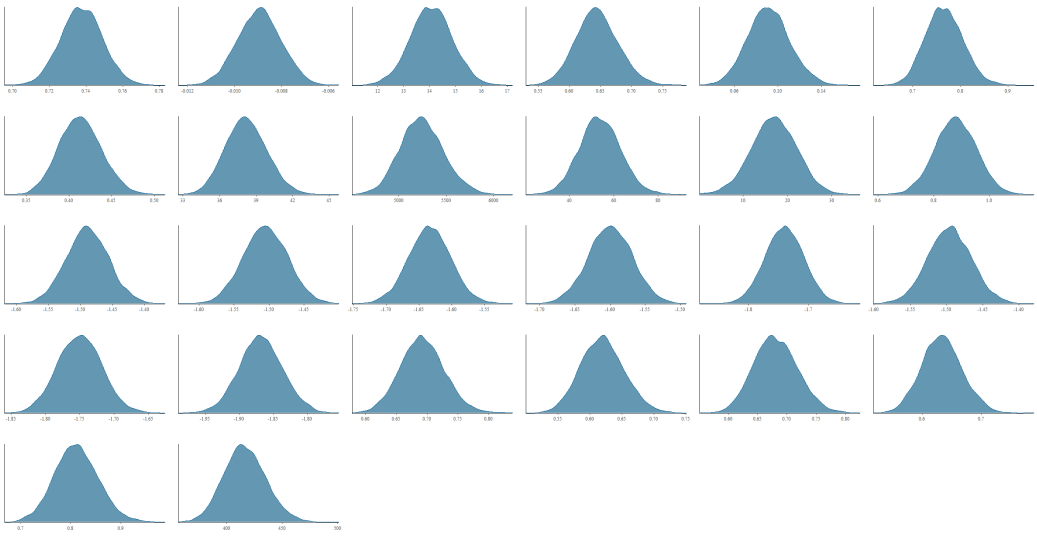
Јаке аутокорелације могу да наговесте проблем са спецификацијом модела, или да укажу на потребу за проређивањем узорка. Ипак, ако је установљена конвергенција, постојање аутокорелација може и да се занемари.



Слика 10: Графици аутокорелација у ланцима МЛМК параметара модела. За неке од параметара уочавамо позитивну, за неке негативну, корелисаност узастопних опсервација, апсолутне вредности од око 0.3.

6. *Смисленост апостериорних расподела*

Да ли постоје генерисане опсервације које излазе из смисленог домена вредности за дати параметар? На пример, да ли домени апостериорних расподела за параметре дисперзија обухватају негативне вредности? Уз то, слично тачки 4, вреди обратити пажњу на информативност и глаткост густина расподела.



Слика 11: Узорачке густине апостериорних расподела параметара су сконцентрисане око средњих вредности, које јесу смислене.

7. *Третирање параметра коваријационе матрице* Φ

Препорука је да се упореде оцене модела са оценама добијеним када се, уместо одвојеног третирања и избора априорних расподела дисперзија и коваријације независних латентних величина, употреби Вишхартова расподела вишедимензионалног параметра Φ (ознака у теоријским поглављима). Због техничких ограничења, овај корак ће бити прескочен. Наиме, библиотека blavaan, у комбинацији са Stan методом за спровођење МЛМК, не подржава постављање априорне расподеле на вишедимензионални параметар Φ.

8. *Употреба неинформативних априорних расподела*

Прави се контраст између оцена модела и оцена добијених када се пође од неинформативних априорних расподела. На тај начин сазнајемо какав је ефекат субјективно изабраних информативних расподела.

За неинформативне расподеле свих слободних параметара стандардне девијације узете су $\mathrm{IG}(6, 10)$. За преостале слободне параметре користимо расподеле $\mathcal{N}(0, 1)$. Испробани експеримент са неинформативним расподелама сумиран је у табели испод (суфикс „неинфо" односи се на новооцењени модел).

Параметар	**Оцена**	**Оцена (неинфо)**
c	-0.009	-0.007
σ^2	0.153	0.150
μ	0.883	1.629
σ_0^2	0.403	0.430
σ_1^2	0.625	0.624
σ_2^2	0.754	0.777
σ_3^2	0.616	0.621
σ_6^2	0.663	0.655
σ_7^2	0.606	0.629
σ_8^2	0.782	0.773
σ_9^2	0.671	0.702
μ_8	-1.493	1.286
μ_1	-1.509	0.405
μ_9	-1.633	1.321
μ_7	-1.599	1.363
μ_2	-1.740	0.285
μ_0	-1.495	0.583
μ_6	-1.746	1.363
μ_3	-1.865	0.307
μ_{IR}	13.694	15.675
λ_{IR}	0.743	0.720
$\mathbf{D}(Gov_{10})$	5341.767	5331.100
$\mathbf{D}(IS_{10})$	61.344	82.993
$\mathbf{D}(IR_{10})$	11.956	3.027
$\mathbf{D}(log\_T_r)$	38.761	13.393
$cov(log\_T_r, Gov_{10})$	423.853	217.760

Упечатљива разлика су оцене одсечака. При неинформативним априорним расподелама, одсечци измерених логаритама бројева позитивних случајева вируса су позитивни. То значи да интерпретација не може да се преточи у смислену мултипликативну форму, на начин хипотетисан у уводу поглавља 4. Такође, одсечак μ је увећан. Интересантно је да се коефицијент c није много променио. Дакле, овај корак у дијагностици подржаваће стабилност закључка о интензитету утицаја мера власти на преносивост вируса (интерпретација следи у засебном одељку).

У контексту тачака 7 и 8, извештавамо резултате још једаног експеримента, у вези са избором априорних расподела. Ради се о идентичном, почетном скупу информативних априорних расподела коефицијента оптерећења, регресионог коефицијента и свих одсечака. Међутим, овога пута користимо аутоматски предложене, неинформативне расподеле пакета blavaan, за све параметаре стандардне девијације, што је гама расподела са параметрима $(1, .5)$. Добијене оцене модела нећемо наводити, уз напомену да ни једна није значајно одступила од оцена у поглављу 4.2.

9. *Осетљивост на избор информативних априорних расподела*

Последњи корак, који доприноси потпуној транспарентности, подразумева испитивање утицаја одабраних хипер-параметара априорних расподела на оцене MCJ. Предлаже се померање хипер-параметара положаја улево и удесно, као и варирање хипер-параметара дисперзије. Најпре представљамо 5 алтернативних избора хипер-параметара априорних расподела (A1, A2, A3, A4, A5), затим дајемо оцене параметара које им одговарају.

Параметар	**Расподела**	**Хиперпар. - A1**	**Хиперпар. - A2**	**Хиперпар. - A3**	**Хиперпар. - A4**	**Хиперпар. - A5**
c	$\mathcal{N}(-0.015, 0.005)$	(-0.05,0.005)	(-0.2,0.005)	(-0.015,0.005)	(-0.015,0.015)	(-0.015,0.015)
σ	$\mathrm{IG}(9, 4)$	(9,4)	(9,4)	(9,4)	(9,4)	(9,4)
μ	$\mathcal{N}(1.125, 0.25)$	(2, 0.25)	((1, 0.25)	(1.125, 0.125)	(1.125, 0.15)	(1.125, 0.15)
$\sigma_0, \sigma_1, \sigma_2, \sigma_3$	$\mathbf{IG}(11, 3)$	(11,3)	(11,3)	(11,3)	(11,3)	(11,3)
$\sigma_6, \sigma_7, \sigma_8, \sigma_9$	$\mathbf{IG}(11, 3)$	(11,3)	(11,3)	(11,3)	(11,3)	(11,3)
μ_8	$\mathcal{N}(-0.95, 0.05)$	(-0.95, 0.1)	(-0.95,0.1)	(-0.95,0.05)	(-0.95,0.1)	(-0.95,0.05)
μ_1	$\mathcal{N}(-0.95, 0.05)$	(-0.95, 0.1)	(-0.95,0.1)	(-0.95,0.05)	(-0.95,0.1)	(-0.95,0.05)
μ_9	$\mathcal{N}(-1.65, 0.05)$	(-0.95, 0.1)	(-1.65,0.1)	(-1.65,0.05)	(-1.65,0.1)	(-1.65,0.05)
μ_7	$\mathcal{N}(-1.65, 0.05)$	(-1.65,0.1)	(-1.65,0.1)	(-1.65,0.05)	(-1.65,0.1)	(-1.65,0.05)
μ_2	$\mathcal{N}(-1.65, 0.05)$	(-1.65,0.1)	(-1.65,0.1)	(-1.65,0.05)	(-1.65,0.1)	(-1.65,0.05)
μ_0	$\mathcal{N}(-1.65, 0.05)$	(-1.65,0.1)	(-1.65,0.1)	(-1.65,0.05)	(-1.65,0.1)	(-1.65,0.05)
μ_6	$\mathcal{N}(-2.35, 0.05)$	(-2.35,0.1)	(-2.35,0.1)	(-2.35,0.05)	(-2.35,0.1)	(-2.35,0.05)
μ_3	$\mathcal{N}(-2.35, 0.05)$	(-2.35,0.1)	(-2.35,0.1)	(-2.35,0.05)	(-2.35,0.1)	(-2.35,0.05)
μ_{IR}	$\mathcal{N}(0, 32)$	(0,32)	(0,32)	(0,32)	(0,32)	(0,32)
λ_{IR}	$\mathcal{N}(0.7, 0.1)$	(0.7,0.1)	(0.8,0.1)	(0.7,0.05)	(0.7,0.15)	(0.7,0.15)
$\sqrt{\mathbf{D}(Gov_{10})}$	$\mathrm{IG}(6, 10)$	(6,10)	(6,10)	(6,10)	(6,10)	(6,10)
$\sqrt{\mathbf{D}(IS_{10})}$	$\mathrm{IG}(6, 10)$	(6,10)	(6,10)	(6,10)	(6,10)	(6,10)
$\sqrt{\mathbf{D}(IR_{10})}$	$\mathrm{IG}(6, 10)$	(6,10)	(6,10)	(6,10)	(6,10)	(6,10)
$\sqrt{\mathbf{D}(log\_T_r)}$	$\mathrm{IG}(6, 10)$	(6,10)	(6,10)	(6,10)	(6,10)	(6,10)
$cov(log\_T_r, Gov_{10})$	$\beta(1, 1)$	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)

Пар.	**Оцена**	**Оцена - A1**	**Оцена - A2**	**Оцена - A3**	**Оцена - A4**	**Оцена - A5**
c	-0.009	-0.011	-0.016	-0.010	-0.009	-0.009
σ^2	0.153	0.158	0.171	0.153	0.156	0.153
μ	0.883	1.178	1.421	0.938	0.989	0.907
σ_0^2	0.403	0.756	0.625	0.402	0.404	0.607
σ_1^2	0.625	0.602	0.626	0.601	0.626	0.607
σ_2^2	0.754	0.756	0.625	0.754	0.754	0.607
σ_3^2	0.616	0.632	0.608	0.616	0.598	0.664
σ_6^2	0.663	0.632	0.634	0.664	0.632	0.664
σ_7^2	0.606	0.613	0.625	0.607	0.609	0.607
σ_8^2	0.782	0.746	0.745	0.782	0.746	0.782
σ_9^2	0.671	0.674	0.674	0.671	0.673	0.626
μ_8	-1.493	-1.501	-1.556	-1.488	-1.515	-1.490
μ_1	-1.509	-1.635	-1.585	-1.514	-1.617	-1.512
μ_9	-1.633	-1.516	-1.572	-1.628	-1.531	-1.629
μ_7	-1.599	-1.475	-1.532	-1.593	-1.490	-1.595
μ_2	-1.740	-1.786	-1.736	-1.746	-1.768	-1.744
μ_0	-1.495	-1.505	-1.453	-1.501	-1.486	-1.499
μ_6	-1.746	-1.517	-1.574	-1.741	-1.532	-1.742
μ_3	-1.865	-1.807	-1.757	-1.871	-1.788	-1.869
μ_{IR}	13.694	13.600	12.969	13.783	13.722	13.663
λ_{IR}	0.743	0.745	0.754	0.742	0.743	0.744
$\mathbf{D}(Gov_{10})$	5341.767	5333.183	5322.459	5341.470	5337.697	5337.045
$\mathbf{D}(IS_{10})$	61.344	63.072	62.627	60.817	62.544	61.500
$\mathbf{D}(IR_{10})$	11.956	10.978	10.803	12.321	11.351	11.820
$\mathbf{D}(log\_T_r)$	38.761	37.378	38.129	38.689	37.548	38.679
$cov(log\_T_r, Gov_{10})$	423.853	415.171	419.923	423.440	416.211	423.179

Померањем хипер-параметара положаја, оцене параметара μ и c се мењају. При различитим комбинацијама априорних расподела, добијају се различито рангирани отсечци $\mu_0, \mu_1, \mu_2, \mu_3$, односно $\mu_6, \mu_7, \mu_8, \mu_9$.

4.4 Анализа осетљивости

Као надоградња на завршетак листе бајесовске дијагностике, наводимо („необавезне”) испробане кораке, који ће пружити детаљнији увид у валидност главног, оцењеног модела из одељка 4.2.

Два проблематична запажања у вези са моделом су:

1. PPP-вредност модела, базирана на χ^2 функцији несагласности, је тачно 0.

2. Дисперзија латентне величине Gov_{10} је огромна.

Који су разлози ових (вероватно међусобно повезаних) неповољних одлика? Можемо да посумњамо у неколико праваца.

4.4.1 Узорковање

Прво, као што смо назначили на почетку поглавља 4.2, инстанце доступних података су очигледно корелисане. Због тога смо случајним избором извукли подскуп обима 1000. Сада ћемо то извлачење реплицирати, како би се омогућила детекција потенцијалне пристрасности која потиче од конкретног узорка.

Пар.	**Оцена**	**Оцена - U1**	**Оцена - U2**	**Оцена - U3**	**Оцена - U4**	**Оцена - U5**
c	-0.009	-0.008	-0.010	-0.007	-0.009	-0.009
σ^2	0.153	0.186	0.099	0.268	0.147	0.095
μ	0.883	0.898	0.934	0.793	0.852	0.823
σ_0^2	0.403	0.452	0.432	0.483	0.423	0.415
σ_1^2	0.625	0.808	1.027	0.875	0.713	0.806
σ_2^2	0.754	0.913	0.669	0.722	0.714	0.592
σ_3^2	0.616	0.8	0.697	0.627	0.714	0.682
σ_6^2	0.663	0.727	0.642	0.784	0.746	0.594
σ_7^2	0.606	0.734	0.728	0.807	0.747	0.588
σ_8^2	0.782	1.023	0.808	0.701	0.799	0.735
σ_9^2	0.671	0.984	0.683	0.874	0.818	0.526
μ_8	-1.493	-1.387	-1.490	-1.471	-1.435	-1.483
μ_1	-1.509	-1.465	-1.487	-1.527	-1.521	-1.544
μ_9	-1.633	-1.598	-1.646	-1.664	-1.657	-1.631
μ_7	-1.599	-1.545	-1.612	-1.595	-1.603	-1.625
μ_2	-1.740	-1.651	-1.725	-1.716	-1.685	-1.694
μ_0	-1.495	-1.439	-1.501	-1.507	-1.510	-1.525
μ_6	-1.746	-1.656	-1.719	-1.729	-1.750	-1.720
μ_3	-1.865	-1.81	-1.895	-1.864	-1.894	-1.848
μ_{IR}	13.694	14.074	11.568	12.932	11.837	12.037
λ_{IR}	0.743	0.757	0.767	0.754	0.772	0.768
$\mathbf{D}(Gov_{10})$	5341.767	5507.251	5211.968	5007.955	5083.508	5001.517
$\mathbf{D}(IS_{10})$	61.344	73.746	62.054	58.370	70.048	72.183
$\mathbf{D}(IR_{10})$	11.956	28.645	11.973	13.085	5.211	3.378
$\mathbf{D}(log\_T_r)$	38.761	39.665	37.312	35.579	35.741	37.461
$cov(log\_T_r, Gov_{10})$	423.853	394.807	406.258	389.827	394.224	399.047

Сви узорци (U1, U2, U3, U4, U5) су прости случајни, обима 1000. Оцене параметара са нормалним априорним расподелама не варирају много при промени узорка, што се не може рећи за параметре дисперзије. Ипак, параметри дисперзије су мање важни при интерпретацији оцењеног модела.

Пар.	**Оцена**	**Оцена - U6**	**Оцена - U7**
c	-0.009	-0.010	-0.005
σ^2	0.153	0.026	0.088
μ	0.883	0.959	0.586
σ_0^2	0.403	0.757	0.753
σ_1^2	0.625	2.302	0.821
σ_2^2	0.754	0.590	0.673
σ_3^2	0.616	0.742	0.961
σ_6^2	0.663	0.810	0.808
σ_7^2	0.606	0.738	0.750
σ_8^2	0.782	1.184	0.873
σ_9^2	0.671	0.778	0.779
μ_8	-1.493	-1.128	-1.389
μ_1	-1.509	-1.097	-1.468
μ_9	-1.633	-1.676	-1.678
μ_7	-1.599	-1.661	-1.586
μ_2	-1.740	-1.727	-1.706
μ_0	-1.495	-1.603	-1.445
μ_6	-1.746	-2.115	-1.850
μ_3	-1.865	-2.181	-2.002
μ_{IR}	13.694	13.645	11.831
λ_{IR}	0.743	0.739	0.763
$\mathbf{D}(Gov_{10})$	5341.767	4309.063	4904.373
$\mathbf{D}(IS_{10})$	61.344	48.772	59.440
$\mathbf{D}(IR_{10})$	11.956	18.641	11.981
$\mathbf{D}(log\_T_r)$	38.761	27.106	30.727
$cov(log\_T_r, Gov_{10})$	423.853	306.220	351.079

Узорци U6 и U7 су стратификовани - за сваку од 146 земаља света извучен је случајно 1, односно 5, датума са свим измереним величинама. На тај начин додатно смањујемо корелисаност између опсервација, али смањујемо и обим узорка. Подаци су ближи претпоставкама метода, али је већи утицај априорних расподела на оцене. Оба процеса МЛМК јесу достигла конвергенцију. Оцене неких параметара се знатно разликују од главног модела, што треба имати у виду приликом интерпретације главног модела.

4.4.2 Симулирани подаци

Други вид одступања од претпоставки ЛМСЈ је то што измерене променљиве нису идеално нормално расподељене. Треће, може се сумњати да се одређена нефлексибилност алгоритмима оцењивања уводи тиме што смо фиксирали велики број параметара, на вредности 0 или 1. Четврто, није извесно да је обим узорка 1000 довољно велик да опише све односе међу величинама, према постављеној спецификацији модела. Да бисмо испитали способност алгоритама ЛМСЈ, са Stan алгоритмом МЛМК, да препозна тачне вредности параметара модела, приступићемо експерименту моделовања симулираних података. Податке симулирамо тако да грубо прате својства узорка који је коришћен за оцењивање главног модела.

Генеришемо 1000 инстанци доле наведених случајних величина, на наведени начин, наведеним редоследом. Називи случајних величина су еквивалентни називима одговарајућих величина које учествују у МСЈ из предходних одељака, са додатим префиксом $S\_$.

$$\mu_1 = \begin{bmatrix} 4.5 \\ 70 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2.5 & 5 \\ 5 & 45 \end{bmatrix}$$

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$$

$$\mu_2 = \begin{bmatrix} 4.5 \\ 82 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2.5 & 2 \\ 2 & 15 \end{bmatrix}$$

$$X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

$$P \sim Ber(0.5)$$

$$\begin{bmatrix} S\_log\_T_r \\ S\_log\_Gov_{10} \end{bmatrix} = P \cdot X_1 + (1 - P) \cdot X_2$$

$$\mu_{11} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma_{11} = \begin{bmatrix} 1 & 0.8 & 0.75 & 0.7 \\ 0.8 & 1 & 0.8 & 0.75 \\ 0.75 & 0.8 & 1 & 0.8 \\ 0.7 & 0.75 & 0.8 & 1 \end{bmatrix}$$

$$E_1 \sim \mathcal{N}(\mu_{11}, \Sigma_{11})$$

$$\begin{bmatrix} S\_log\_N_9 \\ S\_log\_N_8 \\ S\_log\_N_7 \\ S\_log\_N_6 \end{bmatrix} = \begin{bmatrix} log(0.2) \\ log(0.4) \\ log(0.2) \\ log(0.1) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} S\_log\_T_r + E_1$$

$$\mu_{22} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{22} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

$$E_2 \sim \mathcal{N}(\mu_{22}, \Sigma_{22})$$

$$\begin{bmatrix} S\_IS_{10} \\ S\_IR_{10} \end{bmatrix} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix} S\_log\_Gov_{10} + E_2$$

$$E \sim \mathcal{N}(0, 0.5)$$

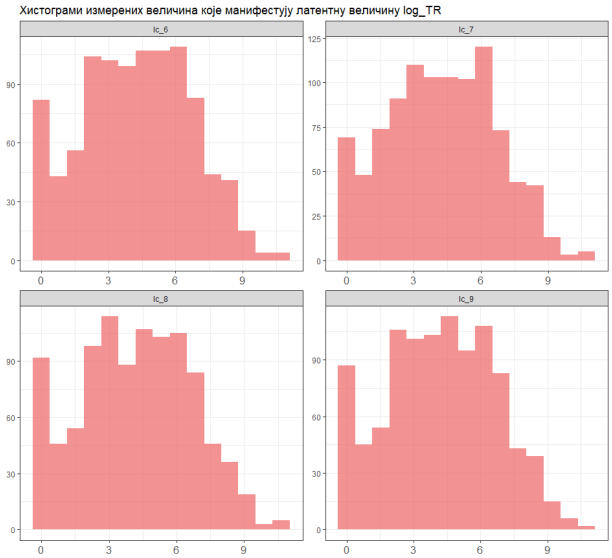
$$S\_log\_T_d = log(3) + S\_log\_T_r + (-0.02) \cdot S\_Gov_{10} + E$$

$$\mu_{33} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma_{33} = \begin{bmatrix} 1 & 0.8 & 0.75 & 0.7 \\ 0.8 & 1 & 0.8 & 0.75 \\ 0.75 & 0.8 & 1 & 0.8 \\ 0.7 & 0.75 & 0.8 & 1 \end{bmatrix}$$

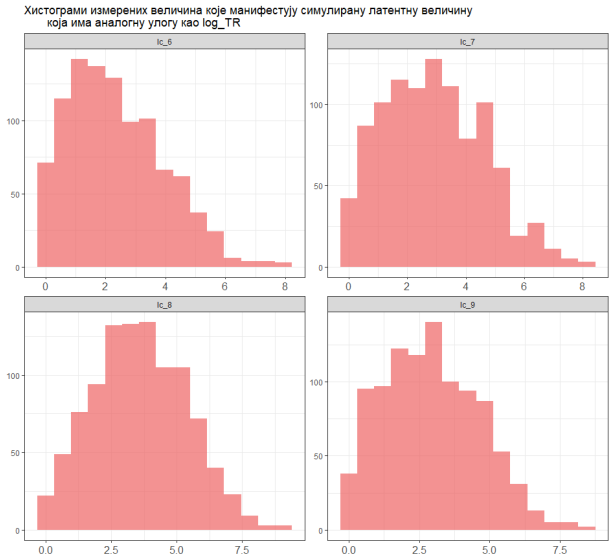
$$E_3 \sim \mathcal{N}(\mu_{33}, \Sigma_{33})$$

$$\begin{bmatrix} S\_log\_N_0 \\ S\_log\_N_1 \\ S\_log\_N_2 \\ S\_log\_N_3 \end{bmatrix} = \begin{bmatrix} log(0.2) \\ log(0.4) \\ log(0.2) \\ log(0.1) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} S\_log\_T_d + E_3$$

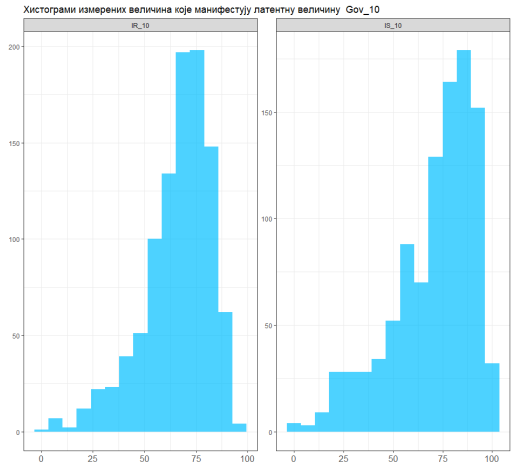
На овај начин, добили смо симулиране измерене величине, чије расподеле представљамо у виду хистограма. Они заиста личе на хистограме стварних података.



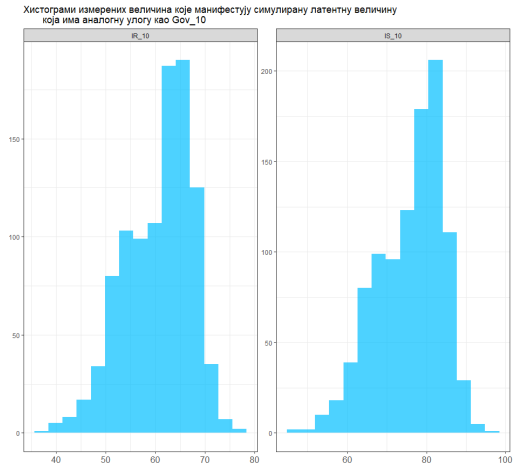
Слика 12: Узорачке расподеле измерених величина које манифестују латентну log_T_r



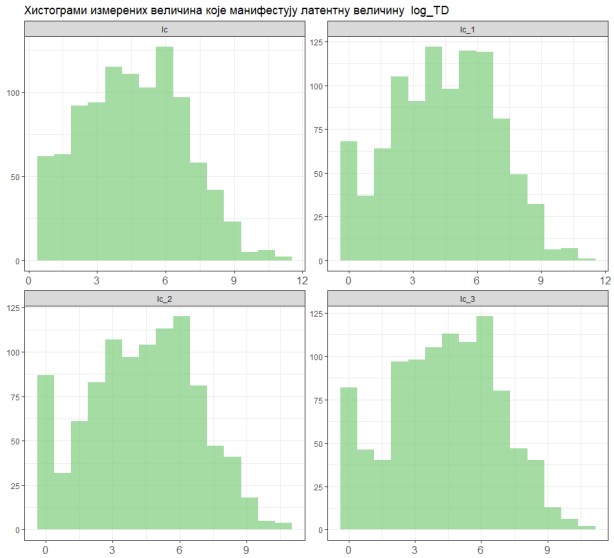
Слика 13: Узорачке расподеле симулираних измерених величина које манифестују латентну $S\_log\_T_r$



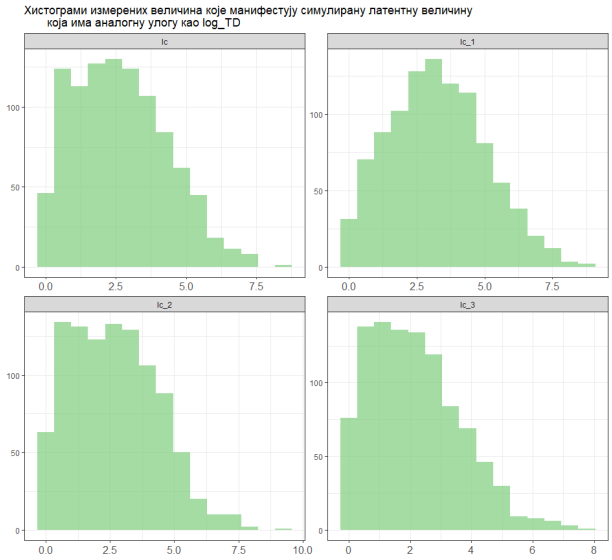
Слика 14: Узорачке расподеле измерених величина које манифестују латентну Gov_{10}



Слика 15: Узорачке расподеле симулираних измерених величина које манифестују латентну S_Gov_{10}



Слика 16: Узорачке расподеле измерених величина које манифестују латентну log_T_d



Слика 17: Узорачке расподеле симулираних измерених величина које манифестују латентну $S\_log\_T_d$

Подаци су симулирани баш тако да им одговара спецификација MCJ са следећим параметрима:

$$\log(T_d) = \log(3) + 1 \cdot \log(T_r) + (-0.02) \cdot Gov_{10} + \delta$$

$$\begin{bmatrix} log(N_9) \\ log(N_8) \\ log(N_7) \\ log(N_6) \\ log(N_0) \\ log(N_1) \\ log(N_2) \\ log(N_3) \\ IS_{10} \\ IR_{10} \end{bmatrix} = \begin{bmatrix} log(0.2) \\ log(0.4) \\ log(0.2) \\ log(0.1) \\ log(0.2) \\ log(0.4) \\ log(0.2) \\ log(0.1) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0.8 \end{bmatrix} \begin{bmatrix} \log(T_r) \\ \log(T_d) \\ Gov_{10} \end{bmatrix} + \epsilon$$

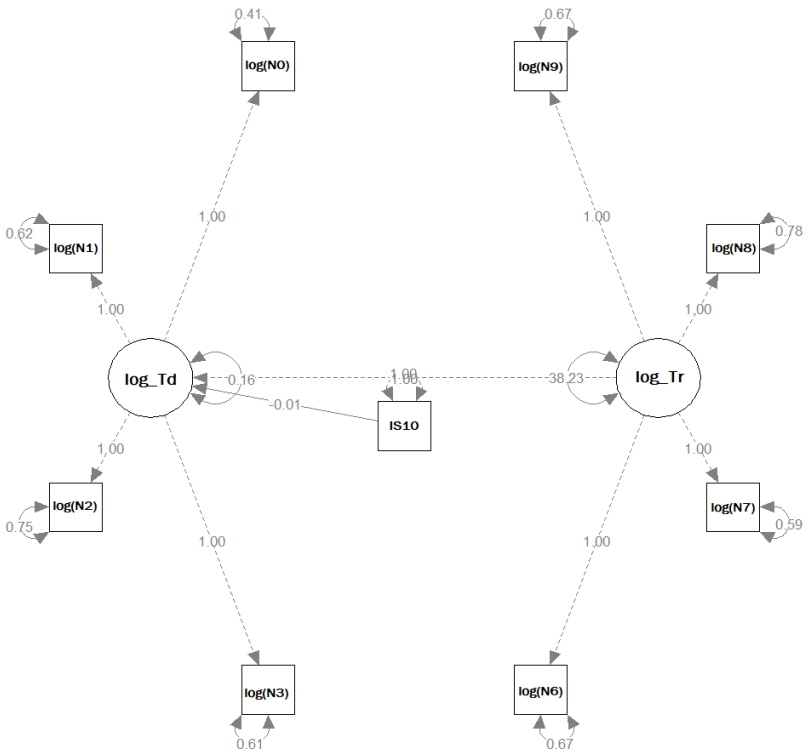
Покушаћемо да оценимо MCJ на симулираним подацима. Избор хипер-параметара МЛМК процеса и априорних расподела извршен је на исти начин као за оцењивање главног модела. Оцене су дате у табели у наставку.

Упркос томе што је спецификација модела тачна, априорне расподеле су густо сконцентрисане око стварних вредности и опсервације јесу независне, алгоритам оцењивања није пронашао исправне вредности свих параметара. Приметимо да је структура оцена коваријационих параметара слична оној код модела оцењеног на стварним подацима. На пример, поново је $\mathbf{D}(Gov_{10})$ огромна. И друге дисперзије одступају од тачних вредности - за модел мерења су подцењене, а за структурни модел прецењене. Ипак, оцене параметара који су кључни за интерпретацију главног модела, c и μ, су у овом примеру блиске тачним вредностима. Такође, и за овај модел је PPP-вредност једнака 0. Све ово узимајући у обзир, закључујемо да дијагностика главног модела није претерано проблематична, па модел може да се користи и интерпретира.

Пар.	**Стварна вредност**	**Оцена**	**Стд. дев.**	**Иннтервал покривања**
c	-0.02	-0.017	0.002	[-0.022,-0.012]
σ^2	0.5	1.448	0.069	[1.319,1.589]
μ	$\log(3) = 1.098$	1.027	0.183	[0.678,1.384]
σ_0^2	1	0.228	0.015	[0.201,0.258]
σ_1^2	1	0.287	0.017	[0.255,0.32]
σ_2^2	1	0.182	0.013	[0.158,0.208]
σ_3^2	1	0.449	0.024	[0.403,0.497]
σ_6^2	1	0.353	0.019	[0.316,0.392]
σ_7^2	1	0.199	0.013	[0.175,0.226]
σ_8^2	1	0.261	0.015	[0.233,0.293]
σ_9^2	1	0.265	0.016	[0.235,0.297]
μ_8	$\log(0.4) = -0.91$	-1.014	0.028	[-1.067,-0.96]
μ_1	$\log(0.4) = -0.91$	-1.064	0.029	[-1.122,-1.01]
μ_9	$\log(0.2) = -1.609$	-1.645	0.028	[-1.699,-1.591]
μ_7	$\log(0.2) = -1.609$	-1.667	0.028	[-1.72,-1.612]
μ_2	$\log(0.2) = -1.609$	-1.669	0.028	[-1.724,-1.614]
μ_0	$\log(0.2) = -1.609$	-1.667	0.028	[-1.723,-1.613]
μ_6	$\log(0.1) = -2.302$	-2.237	0.029	[-2.293,-2.181]
μ_3	$\log(0.1) = -2.302$	-2.205	0.030	[-2.264,-2.147]
μ_{IR}	0	1.167	0.300	[0.573,1.755]
λ_{IR}	0.8	0.785	0.004	[0.777,0.793]
$\mathbf{D}(Gov_{10})$	69.44	5738.429	256.855	[5255.92,6265.992]
$\mathbf{D}(IS_{10})$	1	0.807	0.132	[0.568,1.078]
$\mathbf{D}(IR_{10})$	0.8	0.661	0.084	[0.498,0.822]
$\mathbf{D}(log\_T_r)$	2.58	23.321	1.069	[21.283,25.5]
$cov(log\_T_r, Gov_{10})$	0.91	344.716	16.023	[314.083,377.176]

4.5 Конкурентни модел

Пре тога, покушаћемо да употребимо алернативну спецификацију MCJ. Уместо латентне величине Gov_{10}, као предиктор латентне величине log_T_d послужиће сама променљива IS_{10}, као фиксни предиктор.



Слика 18: Оцене параметара алтернативног MCJ, са фиксним предиктором.

Априорна расподела регресионог коефицијента који се односи на утицај фиксног предиктора IS_{10} на log_T_d биће $\mathcal{N}(-0.015, 0.005)$. Све остале априорне расподеле су одабране идентично, као за главни модел.

Оцењени модел представљамо дијаграмом путање.

Уоредимо два модела на основу информационих критеријума. За главни модел је DIC = 41804.12 и BIC = 41938.14, док алтернативни модел има DIC = 55056.30 и BIC = 55057.36. Апроксимација Бајесовог фактора износи 50627, у корист главног модела. Према томе, главни модел је бољи од алтернативног. Можемо имати у виду и то да метрике за модел оцењен на симулираним подацима износе DIC = 33511.17 и BIC = 33648.19.

4.6 Интерпретација резултата

Решење једначина модела (из 4.2) гласи:

$$\log(T_d) = 0.883 + 1 \cdot \log(T_r) + (-0.009) \cdot Gov_{10} + \log(G)$$

$$\begin{bmatrix} log(N_9) \\ log(N_8) \\ log(N_7) \\ log(N_6) \\ log(N_0) \\ log(N_1) \\ log(N_2) \\ log(N_3) \\ IS_{10} \\ IR_{10} \end{bmatrix} = \begin{bmatrix} -1.633 \\ -1.493 \\ -1.599 \\ -1.746 \\ -1.495 \\ -1.509 \\ -1.740 \\ -1.865 \\ 0 \\ 13.694 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0.743 \end{bmatrix} \begin{bmatrix} \log(T_r) \\ \log(T_d) \\ Gov_{10} \end{bmatrix} + \epsilon$$

Желимо да проценимо вредност R_0.

$$0.883 = \mu = log(R_0) - \frac{\sigma^2}{2} = log(R_0) - \frac{0.153}{2}$$

$$R_0 = \exp(0.9595) = 2.6$$

Дакле, просечан број секундарних случајева заразе који потичу од једног примарног оболелог, у одсуству мера превенције, је 2.6.

Даље, интерпретирамо одсечке из једначина мерења.

$$\mu_i = log(c_i) - \frac{\sigma_i^2}{2}$$

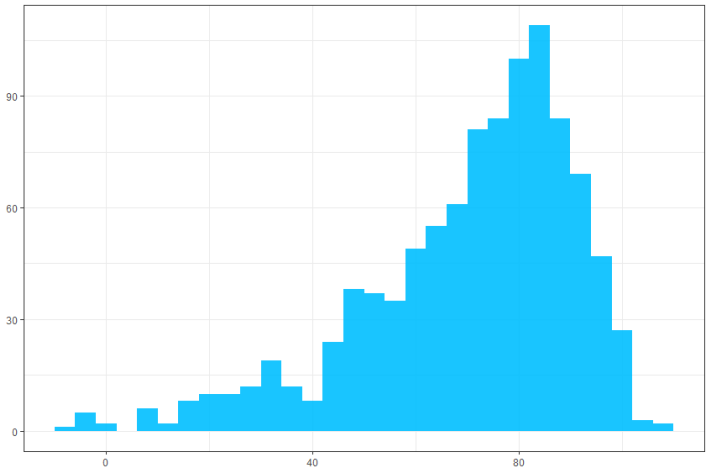
$$c_i = \exp(\mu_i + \frac{\sigma_i^2}{2})$$

$$c_9 = 0.273214, c_8 = 0.332206, c_7 = 0.2736241, c_6 = 0.2430471$$

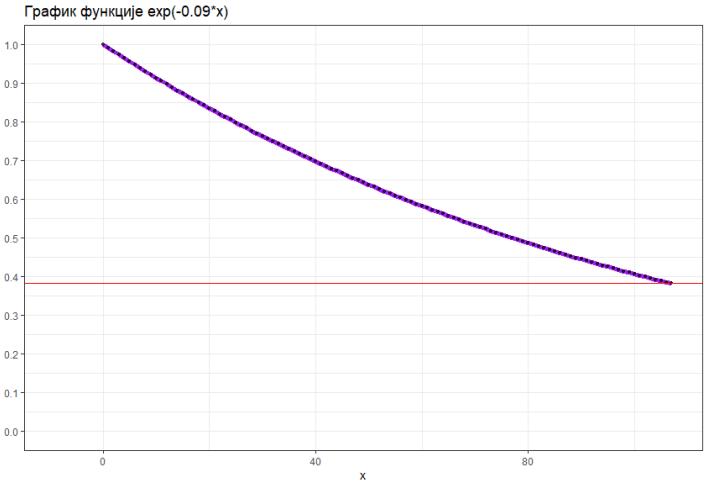
$$c_0 = 0.274309, c_1 = 0.3022502, c_2 = 0.2558919, c_3 = 0.2107674$$

Ово значи да оцењени МСЈ подржава претпоставку да је А највероватније регистрован 8 дана пре референтног датума, а Б 1 дан уназад.

Преостаје да протумачимо оцену параметра $c = -0.009$.
Најпре прикажимо расподелу латентне величине Gov_{10}, на опсервацијама које одговарају параметрима модела генерисаним из апостериорне расподеле параметара током алгоритма МЛМК. Затим на домену вредности латентне величине исцртавамо график функције $\exp(-0.009 \cdot x)$.



Слика 19: Расподела латентне величине Gov_{10}.



Слика 20: График функције $\exp(-0.009 \cdot x)$, на домену величине Gov_{10}.

На крају, на основу изнетог истраживања закључујемо следеће. У одсуству било каквих мера власти, вирус COVID-19 се преноси у складу са репродукционим бројем $R_0 = 2.6$. У случају најоштријих мера, вирус се преноси смањеним интензитетом од 38%, што је у просеку 1 новозаражена особа која потиче од једног примарног случаја.

5 Основни R програм

```
# instaliranje biblioteka
 install.packages("semTools")
 install.packages('semPlot')
 install.packages('rjags')
 install.packages('runjags')
 install.packages('lavaanPlot')
 install.packages("psych")
 remove.packages("rstan")
 if (file.exists(".RData")) file.remove(".RData")
 #Then, restart R.
 install.packages("rstan",repos="https://cloud.r-project.org/",
    dependencies=TRUE)

 if (!require("devtools")) {
   install.packages("devtools")
 }
 devtools::install_github("stan-dev/shinystan",ref="v3-alpha",
    build_vignettes=TRUE)

 devtools::install_github("collectivemedia/tictoc")

 install.packages("lavaan", dependencies=TRUE)
 install.packages("blavaan", dependencies = T)

# ucitavanje biblioteka
library(dplyr)
library(ggplot2)
library(lavaan)
library(blavaan)
library(semPlot)
library(rjags)
#library(shinystan)
library(rstan)
library(tictoc)

# podesavanje paralelnih procesa
```

```
rstan_options(auto_write = TRUE)

future::plan("multiprocess")
options(mc.cores = 3)
Sys.setenv(LOCAL_CPPFLAGS = '-march=corei7 -mtune=corei7')

# n<- parallel::detectCores()/2 # experiment!
# cl <- parallel::makeCluster(4)
# doParallel::registerDoParallel(cl)
# parallel::stopCluster(cl)

# fiksiranje rezultata koji zavise od slucajnih vrednosti
set.seed(1234)

# ucitavanje unapred sredjenih padataka
df <- read.csv('clean_df_rand1000_seed_23.csv')


# specifikacija modela
mod_spec<-
'
#latent variable definition

log_TR=~1*log_new_confirmend_9+1*log_new_confirmend_8+
1*log_new_confirmend_7+1*log_new_confirmend_6

log_TD=~1*log_new_confirmend_3+1*log_new_confirmend_2+
1*log_new_confirmend_1+1*log_new_confirmend

gov_10=~1*IS_10+prior("normal(0.7,0.1)")*IR_10


#regressions

log_TD~1*log_TR+prior("normal(-0.015,0.005)")*gov_10

#intercept

log_TD~prior("normal(1.125,0.25)")*1
```

```
log_new_confirmend_8~prior("normal(-0.95,0.05)")*1
log_new_confirmend_1~prior("normal(-0.95,0.05)")*1

log_new_confirmend_9~prior("normal(-1.65,0.05)")*1
log_new_confirmend_7~prior("normal(-1.65,0.05)")*1
log_new_confirmend_2~prior("normal(-1.65,0.05)")*1
log_new_confirmend~prior("normal(-1.65,0.05)")*1

log_new_confirmend_6~prior("normal(-2.35,0.05)")*1
log_new_confirmend_3~prior("normal(-2.35,0.05)")*1

IS_10~0*1


#vars

log_new_confirmend_8~~prior("gamma(11,3)")*log_new_confirmend_8
log_new_confirmend_1~~prior("gamma(11,3)")*log_new_confirmend_1
log_new_confirmend_9~~prior("gamma(11,3)")*log_new_confirmend_9
log_new_confirmend_7~~prior("gamma(11,3)")*log_new_confirmend_7
log_new_confirmend_2~~prior("gamma(11,3)")*log_new_confirmend_2
log_new_confirmend~~prior("gamma(11,3)")*log_new_confirmend
log_new_confirmend_6~~prior("gamma(11,3)")*log_new_confirmend_6
log_new_confirmend_3~~prior("gamma(11,3)")*log_new_confirmend_3

IS_10~~prior("gamma(6,10)")*IS_10
IR_10~~prior("gamma(6,10)")*IR_10

log_TR~~prior("gamma(6,10)")*log_TR
gov_10~~prior("gamma(6,10)")*gov_10

log_TD~~prior("gamma(9,4)")*log_TD
'


# iscrtavanje dijagrama putanje zarad provere specifikacije
semPaths(lavaan::sem(mod_spec),groups="latents",curvePivot=TRUE,
edge.label.cex=1,label.cex=2)

# fitovanje modela pomocu Stan MCMC, meri se trajanje izvrsavanja
tic("fitting")
```

```
mod_est <- bsem( mod_spec, data=df, bcontrol = list(cores=3),
start='simple', sample = 8000, burnin = 8000, adapt = 2000 ,
n.chains = 3, save.lvs = F)
toc()

# ispisivanje ocena i dijagnostika
summary(mod_est)

fitMeasures(mod_est, 'dic')
fitMeasures(mod_est, 'bic')
blavInspect(mod_est,'neff')

plot(mod_est, plot.type = "acf")
plot(mod_est, plot.type="hist", bins=30)
plot(mod_est, plot.type = "dens")
plot(mod_est, plot.type = "trace")

shinystan::launch_shinystan(as.shinystan(mod_est)) # traje

# sacuvavanje objekta mod_est (fitovanje moze da traje satima)
save(mod_est, file = "mod_est_fin.RData")

# iscrtavanje dijagrama putanje ocenjenog modela
semPaths(mod_est,whatLabels = 'est', pastel=T, groups = "latents",
curvePivot = TRUE)

# druge bajesovske ocene
blavInspect(mod_est,'postmean')
blavInspect(mod_est,'postmedian')
blavInspect(mod_est,'postmode')
```

Литература

[1] Xin-Yuan Song, Sik-Yum Lee, *Basic and Advanced Bayesian Structural Equation Modeling With Applications in the Medical and Behavioral Sciences*, Wiley Series in Probability and Statistics, 2012.

[2] Sjoerd M. H. Huisman, Ahmed Mahfouz, Nematollah K. Batmanghelich, Boudewijn P. F. Lelieveldt, Marcel J. T. Reinders, *A Structural Equation Model for Imaging Genetics Using Spatial Transcriptomics*, 2018.

[3] ESBEN BUDTZ-JØRGENSEN, *Estimation of the Benchmark Dose by Structural Equation Models*, Biostatistics Volume 8, Issue 4, (2007) 675–688.

[4] Joop J. Hox , Timo M. Bechger , *An Introduction to Structural Equation Modelling* , Family Science Review, 11, (1999) 354-373.

[5] Sanne C. Smid, Daniel McNeish, Milica Miočević, Rens van de Schoot, *Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review*, Structural Equation Modeling: A Multidisciplinary Journal, 27:1, (2020) 131-161.

[6] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, *Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models*, Journal of Royal Statistical Society. 64, (1998).

[7] So Yeon Chun, A. Shapiro, *Construction Of Covariance Matrices With A Specified Discrepancy Function Minimizer, With Application To Factor Analysis*, Society for Industrial and Applied Mathematics, J. Matrix Analysis Applications. 31.(2010) 1570-1583.

[8] Edgar C. Merkle, Yves Rosseel, *blavaan: Bayesian Structural Equation Models via Parameter Expansion*, Journal of Statistical Software. 85, (2015).

[9] Marco D'Arienzo, Angela Coniglio, *Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy*, Biosafety and Health, (2020).

[10] Thomas Hale, Noam Angrist, Beatriz Kira, Anna Petherick, Toby Phillips, Thomas Hale, Samuel Webster, *Variation in government responses to COVID-19*, BSG Working Paper Series, 2020.

[11] *Coronavirus disease 2019 (COVID-19) Situation Report – 73*, World Health Organization, 2020.

[12] Odo Diekmann, J.A.P Heesterbeek, Johan Metz, *On the Definition and the Computation of the Basic Reproduction Ratio R0 in Models For Infectious-Diseases in Heterogeneous Populations*, Journal of mathematical biology. 28. (1990) 365-82.

[13] Matthew D. Hoffman, Andrew Gelman, *No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*, Journal of Machine Learning Research 15 (2014) 1593-1623

[14] Sarah Depaoli, Rens van de Schoot, *Improving transparency and replication in Bayesian Statistics: The WAMBS-Checklist*, Psychological Methods. in press., (2016)

Биографија

Милијана Свитлица, рођена 16. октобра 1995. у Ћуприји. Бивши ђак ОШ Бошко Палковљевић Пинки и Земунске гимназије. Уписала програм основних студија Статистика, актуарска и финансијска математика, на Математичком факултету Универзитета у Београду, 2014. године. Наставила са истим програмом мастер студија, 4 године касније. Од 2019. године запослена као члан тима Data Science Services, компаније Etihad Airways, где примењује знања стечена током студија. Међу областима интересовања истиче се биостатистика, што је подстакло избор теме мастер рада.

Захвална породици за подршку током школовања, као и пријатељима и посвећеним наставницима у све три образовне институције. Хвала ментору и комисији за ажурност и конструктивне савете поводом израде мастер рада.