Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет прикладної математики

Проміжний звіт до магістерської дисертації

Математичне та програмне забезпечення системи аналізу та прогнозування макроекономічних даних України на основі емпіричних беєсівських методів

Студент:
Студент групи КМ-31мн,
Цибульник Антон Владиславович

Науковий керівник:
Доцент кафедри ПМА,
Тавров Данило Юрійович

2024

Зміст

1 Формалізація постановки задачі

Об'єкт дослідження: макроекономічні показники під впливом зовнішніх умов; системи аналізу макроекономічних даних; математичні моделі та методи аналізу макроекономічних показників на основі моделей Data Science.

Статистичні методи обробки та аналізу даних; баєсівський підхід до аналізу статистичних даних та емпіричні баєсівські методи серед цього підходу.

Предмет дослідження: емпіричні та ієрархічні беєсівські методи аналізу макроекономічних даних.

Мета дослідження: дослідження та порівняльний аналіз макроекономічних показників під впливом зовнішніх умов на основі емпіричних беєсівських моделей та методів.

Кінцевий результат: система для аналізу макроекономічних даних України, її математичне та програмне забезпечення; оновлені точності побудови статистичних оцінок макроекономічних показників, прогноз їхніх значень.

2 Огляд існуючих рішень за темою магістерської дисертації

2.1 Засоби аналізу макроекономічних даних

2.1.1 Класифікація макроекономічних показників

Макроекономічні показники є важливим інструментом для вивчення економічного розвитку країни та визначення її ефективності у використанні ресурсів. Розглядаючи показники за видами економічної діяльності, можна зробити висновок про внесок різних галузей у формування макроекономічного образу країни.

Система національних рахунків передбачає застосування таких макроекономічних показників, які характеризують результати економічної діяльності країни, зокрема України [1]: випуск, валовий внутрішній продукт, валовий національний дохід, валовий національний наявний дохід та чистий внутрішній продукт.

Випуск (Output)

Випуск відображає обсяг товарів і послуг, вироблених всіма підприємствами в усіх галузях економіки за певний період. Випуск включає валову додану вартість (GVA) — товари й послуги, які купуються для використання, а не для перепродажу чи подальшої обробки — та вартість проміжного споживання, що відображає вартість товарів та послуг, використаних у процесі виробництва.

Валовий внутрішній продукт (GDP)

Валовий внутрішній продукт — це сума всіх кінцевих товарів і послуг, вироблених в економіці за певний період часу, обчислена за вартістю на ринку. Цей показник вимірює первинні доходи лише від внутрішнього виробництва країни, враховує амортизацію або споживання основного капіталу. Валовий внутрішній продукт є центральним макроекономічним

показником країни. Важливість GDP полягає в тому, що він надає комплексну інформацію про рівень економічної активності та ефективність використання ресурсів.

GDP складається з кількох компонентів, включаючи особисте споживання, інвестиції, державні витрати та чистий експорт (експорт мінус імпорт). Оцінюючи ці складові, економісти можуть визначити джерела економічного зростання та виявити можливі проблеми чи ризики.

Один із важливих аспектів GDP — це врахування вартості виробництва, що дозволяє встановити внесок різних галузей економіки. Такий підхід дозволяє визначити, які сектори є основними виробниками, та допомагає сформувати стратегії для підтримки промисловості та сфери послуг.

GDP є не лише показником економічного виробництва, але й індикатором рівня життя. Високий GDP може свідчити про економічний розвиток, але важливо враховувати його розподіл серед населення.

Валовий національний дохід (GNI)

Валовий національний дохід (GNI) є ключовим макроекономічним показником, який враховує економічний внесок країни не лише в її власних межах, але й за їхніми межами. Цей показник обчислюється як сума валового внутрішнього продукту (GDP) та доходів від іноземних інвестицій.

GNI враховує дохід від резидентів країни, які працюють за кордоном, а також дохід іноземних компаній, які працюють в межах країни. Це розширює перспективу оцінки економічної активності та дозволяє отримати повніше уявлення про внесок країни у світову економіку.

Одним з ключових аспектів GNI є його використання для розрахунку національного доходу на душу населення. Цей показник дозволяє визначити розподіл економічного добробуту серед громадян.

Аналіз GNI є важливим для визначення економічної стійкості та здатності країни залучати зовнішні інвестиції. Високий рівень GNI може свідчити про глобальну конкурентоспроможність, водночас його зростання

може слугувати індикатором економічного розвитку.

Валовий національний наявний дохід (GNDI)

Валовий національний наявний дохід (GNDI) являє собою інший аспект оцінки економічної активності країни, що доповнює поняття валового національного доходу (GNI). Як і GNI, GNDI враховує доходи, отримані резидентами країни як в її власних межах, так і за їхніми межами. Проте важливо зрозуміти різницю між цими двома показниками.

Основна відмінність між GNI та GNDI полягає в розрахунках амортизації, що відображає вартість зношення та старіння капіталу у виробництві. GNI містить цей параметр, тобто при визначенні GNI враховується сума витрат на відновлення та заміну виробничого капіталу.

З іншого боку, GNDI є показником «чистого» доходу, який відображається після врахування амортизації. Це означає, що GNDI надає більш точний образ реального доходу, який доступний для витрат чи збереження після усіх обов'язкових витрат, пов'язаних зі зношенням виробничих активів.

Чистий внутрішній продукт (NPD)

Чистий внутрішній продукт (NPD) визначається як валовий внутрішній продукт (GDP) за винятком амортизації, що відображає вартість старіння та зношення виробничого капіталу.

Чистий внутрішній продукт є важливим інструментом для врахування екологічних аспектів економічного розвитку. На відміну від інших макроекономічних показників, NPD дозволяє уникнути перекосів, пов'язаних зі збільшенням виробництва коштом невідповідального використання ресурсів.

Аналізуючи NPD, економісти та урядові структури можуть визначити ефективність економічних стратегій та їхній вплив на сталість виробництва. Цей показник служить також індикатором сталого розвитку, визначаючи, наскільки економічне зростання спрямоване на забезпечен-

ня довгострокового добробуту, ураховуючи аспекти екології та сталості ресурсів.

2.1.2 Інформаційні ресурси макроекономічних даних

Інформаційні ресурси макроекономічних даних в Україні є невідділь-ною частиною економічного аналізу та прийняття урядових та бізнесових рішень. Ці ресурси містять різноманітні статистичні дані, звіти та інші документи, які дають повний образ поточного стану економіки країни.

Національний банк України є ключовим джерелом макроекономічних даних. Він публікує щомісячні та щоквартальні звіти, які включають інформацію про інфляцію, обсяг грошового обігу, обсяги зовнішньої торгівлі та інші важливі економічні показники. Ці дані допомагають визначити тренди у фінансовому секторі та грошовому ринку.

Державна служба статистики України є іншим ключовим джерелом макроекономічних даних. Вона надає інформацію про ВВП, рівень безробіття, виробництво промисловості, сільське господарство та інші соціально-економічні показники. Ці дані важливі для аналізу ефективності різних секторів економіки.

Міністерство економічного розвитку та торгівлі України також надає інформаційні ресурси, зокрема прогнози економічного розвитку, звіти про інвестиції та інші дані, що допомагають приймати урядові рішення щодо економічних стратегій.

Важливим джерелом є також міжнародні організації, такі як Міжнародний валютний фонд (МВФ) та Всесвітній банк, які надають аналітичні звіти та рекомендації з питань економічного розвитку.

Ці інформаційні ресурси використовуються для аналізу економічного стану країни, прогнозування трендів та визначення стратегій для сталого економічного розвитку. Забезпечення доступу до достовірної інформації є важливим чинником для формування ефективних політик та стратегій управління економікою України.

2.1.3 Існуючі системи аналізу економічних даних

Національні банки та статистичні служби

Переважна більшість країн має свої національні банки та служби статистичного аналізу даних, які відповідають за збір, обробку та публікацію макроекономічних даних. Наприклад, в Україні це Національний банк України та Державна служба статистики України.

Міжнародні організації

Організації, такі як Міжнародний валютний фонд (МВФ), Всесвітній банк чи Організація економічного співробітництва та розвитку (ОЕСР), також надають аналітичні звіти та статистичні дані з економічного розвитку країн по всьому світу.

Електронні ресурси

У мережі Інтернет доступні численні електронні ресурси для аналізу та візуалізації економічних даних. До прикладу, це можуть бути вебсайти урядових установ, такі як:

1. Вебсайт Державної служби статистики України [2], підрозділ «Інфографіка» розділу «Доступно про статистику»;
2. Вебсайт Національного банку України [3], розділ «Фінансові ринки»;
3. Вебсайт Міністерства фінансів України [4], підрозділ «Макроекономічний огляд та прогноз» розділу «Дані та статистика».

Іншим прикладом можуть слугувати вебсайти наукових організацій та аналітичних агентств:

1. Інститут економіки та прогнозування НАН України [5];
2. Інститут демографії та соціальних досліджень ім. М. В. Птухи НАН України [6];

3. Інститут регіональних досліджень ім. М. І. Долішнього НАН України [7].

Спеціалізовані програмні продукти

Нижче наведено перелік програмних продуктів та інструментів для аналізу економічних даних, зокрема статистичні пакети та бізнес-аналітичні програми:

1. SAS [8] — це провідний програмний продукт для аналізу даних, який використовується в корпоративному, урядовому та науковому секторах. Він пропонує широкий спектр інструментів для аналізу статистичних даних, прогнозування, моделювання та візуалізації;
2. IBM SPSS Statistics [9] — ще один популярний програмний продукт для аналізу даних;
3. Stata [10] — це потужний програмний продукт для аналізу даних, який часто використовується в академічних дослідженнях.

2.2 Математичні інструменти аналізу макроекономічних даних

2.2.1 Методи регресійного аналізу та часових рядів

Перш за все, наведемо короткий перелік математичних інструментів аналізу макроекономічних даних за допомогою регресійного аналізу та часових рядів [11].

Регресійний аналіз — це статистичний метод, який використовується для дослідження взаємозв'язку між двома або більше змінними. Він дозволяє прогнозувати значення однієї змінної на основі значення іншої або інших змінних. Регресійний аналіз широко використовується в економічних дослідженнях для аналізу макроекономічних даних. Він може використовуватися для таких цілей:

1. Аналіз трендів дозволяє визначити загальний напрямок та зміну розвитку економічного показника;

2. Прогнозування: наприклад, регресійний аналіз може використовуватися для прогнозування темпів зростання ВВП, рівня інфляції або рівня безробіття;

3. Результати регресійного аналізу можуть використовуватися для розробки таких заходів економічної політики, як фіскальна політика, монетарна політика або зовнішня політика.

4. Авторегресійні моделі (AR) — це один із видів регресійних моделей, які використовуються для аналізу часових рядів. AR-моделі визначають значення економічного показника на основі його власних попередніх значень.

5. Модель ARMA (Autoregressive Moving Average) — це модель, яка поєднує в собі autoregressive (AR) і moving average (MA) компоненти. ARMA-моделі можуть бути використані для аналізу часових рядів, які мають як трендову складову, так і випадкову (шумову) складову.

6. Модель ARIMA (Autoregressive Integrated Moving Average) — це модель, яка поєднує в собі autoregressive (AR), moving average (MA) і integrated (I) компоненти. I-компонент відповідає за інтегрування часового ряду, тобто за видалення трендової складової.

7. ARIMA-моделі можуть бути використані для аналізу часових рядів, які мають істотну трендову складову. Наприклад, ARIMA-модель може бути використана для прогнозування темпів зростання ВВП на основі темпів зростання ВВП у попередні періоди з урахуванням трендової складової.

2.2.2 Баєсівський та частотний статистичні підходи до обробки даних

Серед існуючих теоретичних підходів, які використовуються для побудови оцінок показників різних математичних моделей на основі статистичних даних відрізняють так звані частотний та баєсівський підходи.

Частотний підхід у статистичних методах обробки даних трактує невідомий параметр як невідому, але фіксовану величну. Акцент при цьому ставиться на повторюваності подій та довгострокових частотах випадкових явищ, які описують невідомий параметр.

Цей підхід виступає в опозиції до баєсівського підходу, ідея якого полягає у тому, щоб оцінити деякий невідомий параметр певної моделі (наприклад, економічної моделі), вважаючи цей параметр випадковою величиною.

Спираючись на апріорний розподіл параметра (певні початкові уявлення про випадковий характер невідомої величини), а також на наявні статистичні дані, беєсівські методи мають на меті віднайти апостеріорний розподіл параметра, тобто розподіл параметра за умови спостереження наявних даних. У Табл. 1 наведено порівняння частотного та баєсівського підходів.

Таблиця 1: Порівняння частотного та баєсівського підходів

	Частотний підхід	Баєсівський підхід
Тлумачення параметра	Фіксована величина	Випадкова величина
Інформація на вході	Статистичні дані	Статистичні дані та експертні знання (виражених в апріорному розподілі)
Тлумачення результатів	На основі заданої вибірки формуються висновки про значення параметрів генеральної сукупності даних	Результати стосуються оцінок параметрів конкретного набору даних

Продовження на наступній сторінці

Таблиця 1: Порівняння частотного та баєсівського підходів (продовження)

Побудова оцінок параметрів	Точкові статистичні оцінки з довірчими інтервалами	Апостеріорний розподіл, що відображає невизначеність та оновлені знання про параметр
Розмір вибірки даних	Чутливий до розміру вибірки	Менш чутливий, доцільний з невеликими вибірками

Таким чином, особливо у складних моделях, частотний підхід є менш гнучким порівняно з баєсівський підходом, який дозволяє враховувати додаткові експертні знання та оновлювати ймовірності на основі нових даних.

На практиці, при використанні баєсівських методів виникають два питання: яким чином слід обирати апріорний розподіл та що робити у випадку, якщо апостеріорний розподіл має складний для аналізу аналітичний вигляд.

Друге питання вирішується завдяки алгоритмам побудови марковських ланцюгів методом Монте-Карло. Наступні ж підрозділи стосуватимуться огляду існуючих практичних рішень [12] при виборі апріорного розподілу у баєсівському підході.

Informative Priors

Розподіли, які включають конкретні апріорні знання та переконання про параметри моделі, називають інформативними апріорними розподілами (Informative Priors). Іншими словами, інформативні апріорні розподіли містять інформацію про те, які значення параметрів є більш або менш ймовірними, виходячи з попереднього знання або досвіду.

Інформативні розподіли можуть допомогти покращити точність і надійність результатів Баєсівського аналізу. Однак, якщо апріорний роз-

поділ є занадто інформативним, він може призвести до суб'єктивності результатів аналізу. Отже, такі розподіли варто використовувати лише при глибокому розумінні проблеми та у ситуаціях, де експертне знання не є обмеженим.

Conjugate Priors

Апріорний розподіл є спряженим (Conjugate Prior) для певної сім'ї розподілів, якщо апріорний та апостеріорний розподіли належать до тієї ж самої сім'ї розподілів.

Спряжені розподіли корисні тим, що ми завжди отримуємо апостеріорні розподіли в аналітичній формі. В цьому і полягає математична зручність їхнього використання.

До прикладу, наведемо деякі поширені пари спряжених розподілів: спряжений до експоненціальної сім'ї буде також експоненціальний розподіл; нормальний розподіл матиме нормальний розподіл в якості спряженого; спряженим до біноміального розподілу буде Бета-розподіл, спряженим до розподілу Пуассона — Гамма-розподіл тощо.

Однак, на практиці існують численні випадки, коли модель не можна формалізувати певною парою загальновідомих розподілів.

Non-informative Priors

Неінформативні / невпливові апріорні розподіли (Non-informative Priors) є концепцією апріорних розподілів, які передбачають мінімальний виплив переконань дослідника, щоб надати простоту та об'єктивність у баєсівському аналізі.

Іншими словами, основна ідея застосування цих розподілів полягає у тому, щоб обирати апріорний розподіл, який не має значущого впливу на аналіз результатів у випадку, наприклад. обмежених експертних знань про параметри моделі. Це особливо важливо у ситуаціях, коли об'єктивність є пріоритетом, і дослідник хоче, щоб апріорний розподіл не вніс суттєвого впливу на кінцеві результати. Нижче наведено деякі приклади

неінформативних розподілів, таких як Diffuse Priors та Jeffreys' Prior.

Diffuse Priors

Пласкі апріорні розподіли (Diffuse Priors) використовуються тоді, коли дослідник має дуже обмежене або жодне попереднє знання про параметри моделі та прагне досягти максимальної об'єктивності у вивченні даних. Пласкі апріорні розподіли мають великий розкид, тобто рівномірну щільність по всьому простору значень параметра. Прикладом може слугувати нормальний розподіл із нульовим математичним сподіванням та, умовно, нескінченною дисперсією. Такі розподіли дозволяють наявним статистичним даним безпосередньо впливати на вигляд апостеріорного розподілу.

Jeffreys' Prior

Цей невпливовий апріорний розподіл відображає нейтральність щодо параметрів моделі за рахунок властивості інваріантності відносно параметризації. Апріорний розподіл Гарольда Джеффрісона (Jeffreys' Prior) слугує загальним правилом для вибору Non-informative: розподіл обирається пропорційно квадратному кореню з визначника інформаційної матриці Фішера.

Empirical Prior

Водночас емпіричні баєсівські методи замість того, щоб жорстко обирати певний апріорний розподіл, полягають у налаштуванні форми апріорного розподілу на основі вхідних даних.

Іншими словами, ідея полягає в тому, щоб уточнити гіперпараметри розподілу певного показника шляхом аналізу наявних даних, тобто, фактично, використати дані двічі — перший раз для апріорної оцінки параметрів, а другий раз — для оновлення й виведення апостеріорного розподілу.

Такий гібридний підхід дозволяє отримати у певних випадках більш

оптимальні результати. Таким чином, наступні підрозділи охоплюватимуть специфіку, методи та підходи саме емпіричних баєсівських методів.

Висновки до розділу

Серед існуючих математичних інструментів аналізу макроекономічних даних було виділено методи на основі регресійного аналізу та часових рядів, а також статистичні методи баєсівського підходу. Своєю чергою, було продемонстровано, що баєсівський підхід має численні альтернативи для математичної інтерпретації апріорних знань дослідника.

Для розв'язання поставленої в рамках магістерської дисертації задачі, в якості найбільш ефективного методу було обрано групу баєсівських методів, а вже серед них — емпіричні баєсівські методи.

3 Модель системи

Наразі триває пропрацювання можливих математичних інструментів (наступний розділ) для опису моделі системи. Щойно інструменти будуть цілісно опановані, цей розділ набуде свого завершального вигляду.

4 Математичне забезпечення компонентів системи

У цьому розділі детальніше розглянемо ієрархічну Баєсову модель, її особливості використання та застосування. Нехай задано Баєсову модель такого вигляду:

$$X_{ij} \mid \theta_i \overset{\text{ind}}{\sim} \text{Poiss}(\theta_i),\ i = \overline{1,n},\ j = \overline{1,m}, \tag{1}$$

$$\theta_i \mid \sigma \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma),\ i = \overline{1,n}, \tag{2}$$

$$\sigma^{-1} \sim \text{Exp}(1), \tag{3}$$

при цьому k є відомим. Іншими словами, для кожної згенерованої випадкової величини σ з експоненціального розподілу генерується значення θ_i з Гамма-розподілу, і надалі — набір випадкових величин $X_{i1}, \ldots, X_{im}$ з розподілу Пуасона з параметром θ_i.

4.1 Gibbs Sampler

Завдання полягає у тому, щоб навести повні викладки одного кроку вибірки Гіббса для генерування апостеріорних розподілів параметрів θ_i та σ. Однак, перш ніж переходити до алгоритму, наведемо явний вигляд розподілів, які розглядатимуться у подальших міркуваннях.

Дискретна випадкова величина ξ має розподіл Пуасона з параметром λ, якщо:

$$\xi \sim \text{Poiss}(\lambda) \iff P(\xi = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1}$$

Неперервна випадкова величина η має Гамма-розподіл з «shape parameter» $k > 0$ та «scale parameter» $\sigma > 0$, якщо:

$$\eta \sim \text{Gamma}(k, \sigma) \iff f_\eta(x) = \frac{x^{k-1}}{\sigma^k \Gamma(k)}\, e^{-x/\sigma}\, \mathbb{1}(x > 0) \tag{2}$$

Кажуть, що неперервна випадкова величина η^{-1} має обернений

Гамма-розподіл з параметрами $k > 0$ та $\sigma > 0$, якщо:

$$\eta^{-1} \sim \mathrm{IG}(k, \sigma) \iff f_{\eta^{-1}}(x) = \frac{x^{-k-1}}{\Gamma(k)} \sigma^k e^{-\sigma/x} \mathbb{1}(x > 0) \tag{3}$$

Наостанок наведемо вигляд щільности експоненціального розподілу з параметром λ для неперервної випадкової величини ζ^{-1} :

$$\zeta^{-1} \sim \mathrm{Exp}(\lambda) \iff f_{\zeta^{-1}}(x) = \lambda e^{-\lambda x} \mathbb{1}(x > 0), \tag{4}$$

при цьому випадкова величина ζ матиме щільність виду

$$f_{\zeta}(x) = \lambda x^{-2} e^{-\lambda/x} \mathbb{1}(x > 0) \tag{5}$$

Тож нехай ініційовано пару значень $\left(\sigma^{(0)}; \theta_1^{(0)}, \ldots, \theta_n^{(0)}\right)$. Тоді один крок вибірки Гіббса складатиметься із таких пунктів:

1. Використовуючи значення $\sigma^{(0)}$, згенерувати значення $\theta_i^{(1)}$ з так званого «full conditional distribution for parameter θ_i»;
2. Маючи значення $\theta_i^{(1)}$, згенерувати значення $\sigma^{(1)}$ з «full conditional distribution for parameter σ»;

Зауважимо, що повний умовний розподіл одного параметра пропорційний сумісному розроділу, в якому інший параметр вважається фіксованим. Продемонструємо це, застосувавши ланцюгове правило:

$$\underbrace{f(\theta_i, \sigma \mid X_{i1}, \ldots, X_{im})}_{\text{сумісний розподіл}} = \underbrace{f(\theta_i \mid \sigma; X_{i1}, \ldots, X_{im})}_{\text{повний умовний розподіл}} \underbrace{f(\sigma \mid X_{i1}, \ldots, X_{im})}_{\text{не залежить від } \theta_i} \tag{6}$$

Таким чином переконуємося, що повний умовний розподіл як фунція від θ_i пропорційний сумісному розподілу, в якому значення σ покладено фіксованим. Аналогічним чином для параметра σ:

$$\underbrace{f(\theta_i, \sigma \mid X_{i1}, \ldots, X_{im})}_{\text{сумісний розподіл}} = \underbrace{f(\sigma \mid \theta_i; X_{i1}, \ldots, X_{im})}_{\text{повний умовний розподіл}} \underbrace{f(\theta_i \mid X_{i1}, \ldots, X_{im})}_{\text{не залежить від } \sigma} \tag{7}$$

Отже, опишемо один крок вибірки Гіббса:

1. Маючи $\sigma^{(0)}$, згенеруємо значення $\theta_i^{(1)}$ із відповідного умовного розподілу. Послідовно скористаємося такими викладками: щойно продемонтрованою властивістю (6) у переході (1), формулою Баєса у переході (2), ланцюговим правилом у переході (3) та незалежністю значень вибірки даних у кроці (4):

$$\begin{aligned} f(\theta_i \mid \sigma^{(0)}; X_{i1}, \ldots, X_{im}) &\overset{(1)}{\propto} f(\theta_i, \sigma^{(0)} \mid X_{i1}, \ldots, X_{im}) \propto \\ &\overset{(2)}{\propto} f(X_{i1}, \ldots, X_{im} \mid \theta_i, \sigma^{(0)}) f(\theta_i, \sigma^{(0)}) = \\ &\overset{(3)}{=} f(X_{i1}, \ldots, X_{im} \mid \theta_i, \sigma^{(0)}) f(\theta_i \mid \sigma^{(0)}) f(\sigma^{(0)}) = \\ &\overset{(4)}{=} \prod_{j=1}^{m} f(X_{ij} \mid \theta_i, \sigma^{(0)}) f(\theta_i \mid \sigma^{(0)}) f(\sigma^{(0)}) \end{aligned} \tag{8}$$

   З огляду на вигляд функцій щільностей заданої Баєсової моделі (1) – (3), матимемо:

$$\begin{aligned} f(\theta_i \mid \sigma^{(0)}; X_{i1}, \ldots, X_{im}) \propto \prod_{j=1}^{m} \frac{\theta_i^{X_{ij}} e^{-\theta_i}}{X_{ij}!} \times \frac{\theta_i^{k-1}}{\left[\sigma^{(0)}\right]^k \Gamma(k)} e^{-\theta_i/\sigma^{(0)}} \mathbb{1}(\theta_i > 0) \times \\ \times \left[\sigma^{(0)}\right]^{-2} e^{-1/\sigma^{(0)}} \mathbb{1}(\sigma^{(0)} > 0) \end{aligned} \tag{9}$$

   Відкидаючи множники, які не мають фунцкіональної залежності від θ_i, отримуємо такий вираз:

$$f(\theta_i \mid \sigma^{(0)}; X_{i1}, \ldots, X_{im}) \propto \theta_i^{m\overline{X}_i + k - 1} e^{-\theta_i (m + 1/\sigma^{(0)})} \mathbb{1}(\theta_i > 0) \tag{10}$$

   А відтак, аналогічно до перетворень у формулі (**??**) з'ясовуємо вигляд апостеріорного розподілу параметра θ_i :

$$\theta_i \mid \sigma^{(0)}; X_{i1}, \ldots, X_{im} \sim \text{Gamma}\left(k + m\overline{X}_i, \frac{\sigma^{(0)}}{1 + \sigma^{(0)} m}\right) \tag{11}$$

   Отже, використовуючи $\sigma^{(0)}$, генеруємо значення $\theta_i^{(1)}$ із розподілу (11).

2. Маючи $\theta_i^{(1)}$, згенеруємо значення $\sigma^{(1)}$ із відповідного повного умовного розподілу. В силу аналогічних кроків, як це показано для виведення

формули (8), матимемо:

$$f(\sigma \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm}) \propto$$
$$\propto f(X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm} \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; \sigma) f(\theta_1^{(1)}, \ldots, \theta_n^{(1)} \,|\, \sigma) f(\sigma), \tag{12}$$

а отже, зважаючи на незалежність випадкових величин $\theta_1^{(1)}, \ldots, \theta_n^{(1)}$, отримаємо

$$f(\sigma \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm}) \propto$$
$$\propto \prod_{i=1}^{n} f(X_{i1}, \ldots, X_{im} \,|\, \theta_i^{(1)}, \sigma) \prod_{i=1}^{n} f(\theta_i^{(1)} \,|\, \sigma) f(\sigma), \tag{13}$$

і як наслідок властивості вибірки даних:

$$f(\sigma \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm}) \propto$$
$$\propto \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij} \,|\, \theta_i^{(1)}, \sigma) \prod_{i=1}^{n} f(\theta_i^{(1)} \,|\, \sigma) f(\sigma) \tag{14}$$

Відтак, з огляду на вигляд функцій щільностей заданої Баєсової моделі, вираз розписуватиметься як:

$$f(\sigma \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm}) \propto$$
$$\propto \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\left[\theta_i^{(1)}\right]^{X_{ij}} e^{-\theta_i^{(1)}}}{X_{ij}!} \times \prod_{i=1}^{n} \frac{\left[\theta_i^{(1)}\right]^{k-1}}{\sigma^k \Gamma(k)} e^{-\theta_i^{(1)}/\sigma} \mathbb{1}(\theta_i^{(1)} > 0) \times$$
$$\times \sigma^{-2} e^{-1/\sigma} \mathbb{1}(\sigma > 0) \tag{15}$$

Відкидаючи множники, які не мають фунцкіональної залежності від σ, отримуємо такий вираз:

$$f(\sigma \,|\, \theta_1^{(1)}, \ldots, \theta_n^{(1)}; X_{11}, \ldots, X_{1m}; \ldots; X_{n1}, \ldots, X_{nm}) \propto$$
$$\propto \sigma^{-nk-2} \, e^{-(n\overline{\theta^{(1)}}+1)/\sigma} \, \mathbb{1}(\sigma > 0) \tag{16}$$

У виразі, наведеному вище, впізнаємо обернений Гамма-розподіл (3):

$$\sigma \,|\, \theta_i^{(1)}; X_{11}, \ldots, X_{nm} \sim \text{IG}\left(nk + 1, n\overline{\theta^{(1)}} + 1\right) \tag{17}$$

Отже, використовуючи набір $\theta_i^{(1)}$, генеруємо значення $\sigma^{(1)}$ із розподілу (17).

4.2 `R-base` Gibbs Sampler

Виконаємо засобами мови `R` вибірку Гіббса згідно із викладками, наведеними у попередньому підрозділі. Вибірку даних (1) задамо на навчальному прикладі даних. Покладемо $n = k = 3$ та $m = 100$, тобто матимемо Баєсову модель вигляду

$$X_{ij} \,|\, \theta_i \overset{\text{ind}}{\sim} \text{Poiss}(\theta_i), \;\; i = \overline{1,3}, \;\; j = \overline{1,100}, \tag{1}$$

$$\theta_i \,|\, \sigma \overset{\text{i.i.d.}}{\sim} \text{Gamma}(3, \sigma), \;\; i = \overline{1,3}, \tag{2}$$

$$\sigma^{-1} \sim \text{Exp}(1) \tag{3}$$

Тож параметрами вибірки слугуватимуть значення σ та вектор параметрів $(\theta_1, \theta_2, \theta_3)$. Поклавши початкове значення $\sigma^{(0)} = 1$, проведемо 100 кроків для «розгону» алгоритму і фіксуватимемо наступні $N = 10\,000$ кроків для визначення апостеріорного розподілу невідомих параметрів.

Наведемо нижче основні програмні блоки. До прикладу, у Лістингу 1 продемонстровано функцію для оновлення параметрів $(\theta_1, \theta_2, \theta_3)$.

Лістинг 1: Крок оновлення вектора параметрів $(\theta_1, \theta_2, \theta_3)$

```
update_theta <- function(x, sigma, k, n, m) {
    theta_values <- array(dim = n)

    for (i in 1:n) {
        shape_theta <- k + m * mean(x[, i])
        scale_theta <- sigma / (1.0 + sigma * m)

        theta_values[i] <- rgamma(1, shape = shape_theta, scale = scale_theta)
    }

    return(theta_values)
```

```
}
```

Нижче у Лістингу 2 продемонстровано програмну реалізацію оновлення значення σ.

Лістинг 2: Крок оновлення параметра σ

```
update_sigma <- function(x, theta, k, n, m) {
    shape_sigma <- n * k + 1
    scale_sigma <- n * mean(theta) + 1

    sigma_value <- 1 / rgamma(1, shape = shape_sigma, scale = 1 / scale_sigma)

    return(sigma_value)
}
```

Запуск основного алгоритму вибірки Гіббса (Лістинг 3) показано нижче у Лістингу 4.

Лістинг 3: Імплементація алгоритму вибірки Гіббса

```
gibbs <- function(data, n_iter, init, prior, n, m) {
    # Initialize arrays to store variables
    sigma_out <- array(dim = c(n_iter, 1))
    theta_out <- array(dim = c(n_iter, 3))

    sigma_now <- init$sigma

    # Gibbs sampler
    for (i in 1:n_iter) {
        theta_now <- update_theta(
            x = data, sigma = sigma_now, k = prior$k, n = n, m = m
        )
        sigma_now <- update_sigma(
            x = data, theta = theta_now, k = prior$k, n = n, m = m
        )

        theta_out[i, ] <- theta_now
        sigma_out[i, ] <- sigma_now
    }

    cbind(sigma = sigma_out[, 1], theta1 = theta_out[, 1], theta2 = theta_out[,
                                        2], theta3 = theta_out[, 3])
}
```

Лістинг 4: Запуск вибірки Гіббса

```r
# Load the dataset
x <- read.csv("hw5.csv")

# Set main parameters
n <- ncol(x)
m <- nrow(x)

prior <- list()
prior$k <- 3

init <- list()
init$sigma <- 1.0

set.seed(53)
posterior <- gibbs(
    data = x,
    n_iter = 10e3 + 100,
    init = init,
    prior = prior,
    n = n,
    m = m
)

# Exclude first 100 iterations (burn-in period)
posterior <- tail(posterior, -100)
```

На Рис. 1 продемонстровано результати: «trace plot» отриманого ланцюга Маркова та апостеріорний розподіл параметрів σ та θ_1. Перш ніж перейти до подальших висновків, проведемо огляд автоколеляції утвореного ланцюга. Автоколеляція як така вказує на міру лінійної залежності (в межах $[-1, 1]$) поточного значення відносно попередніх значень (lags). З Рис. 2 бачимо, що автокореляція значень практично відсутня.

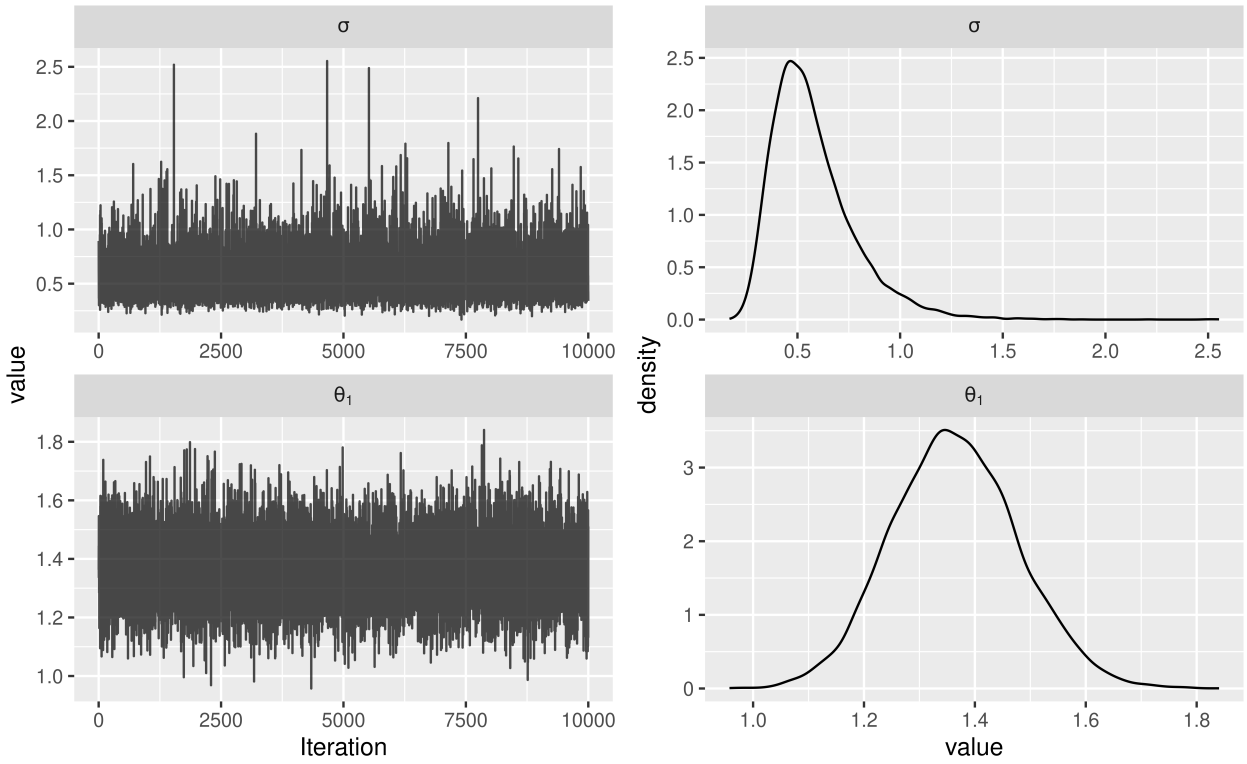

Рис. 1: Апостеріорний розподіл параметрів σ та θ_1 (засоби `R-base`)

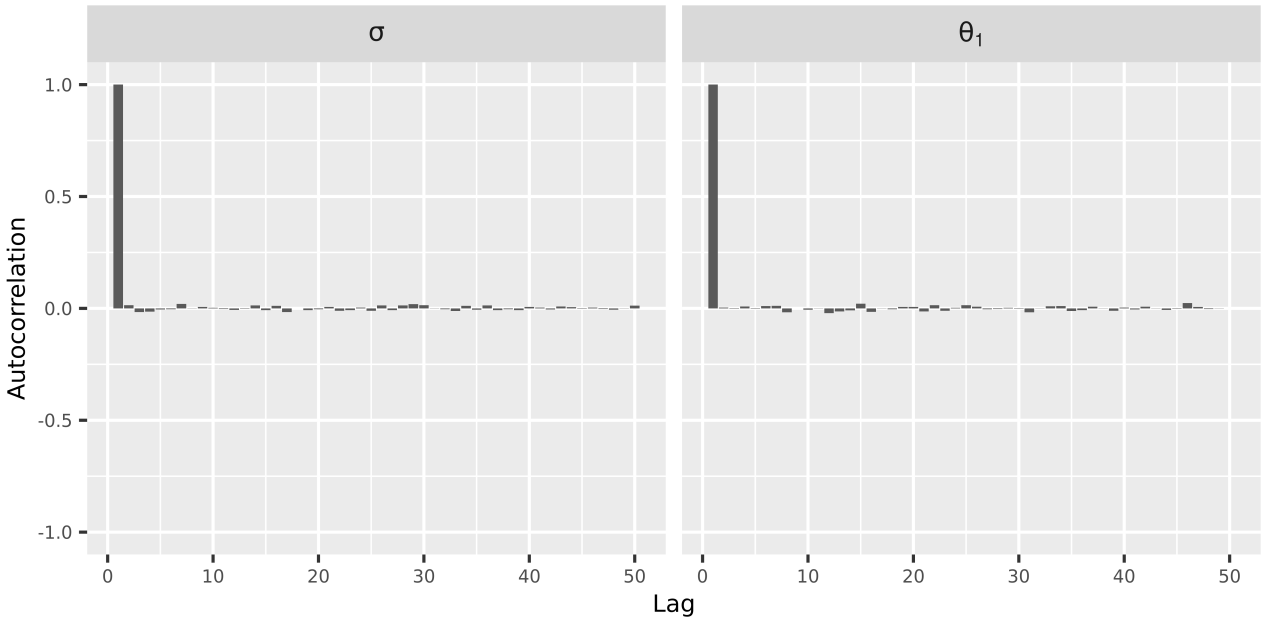

Рис. 2: Автокореляційний графік параметрів σ та θ_1 (засоби `R-base`)

Аналогічні висновки про слабку колеляцію можна зробити із табли-

чного вигляду автоколеляційного графіка (Табл. 2).

	σ	θ_1
Lag 0	1.0000	1.0000
Lag 1	0.0141	0.0038
Lag 5	-0.0036	0.0099
Lag 10	-0.0027	-0.0009
Lag 50	-0.0018	0.0036

Таблиця 2: Автокореляційна таблиця параметрів σ та θ_1

У випадку ж сильної автокореляції значень у пригоді може стати виявлення так званого ефективного розміру вибірки (effective size), який вказуватиме, скільки значень у ланцюзі безпочередньо містять інформацію про стаціонарний розподіл. До прикладу, якщо згенеровано ланцюг довжиною $N = 10\,000$, а ефективний розмір вибірки складає $N_{e.s.} = 1000$, то це означає, що при обчисленні статистичних оцінок про істинний стаціонарний розподіл кожне $N/N_{e.s.} = 100$-те значення містить цінну інформацію, а інші значення є високо корельованими.

Для ланцюга, отриманого на Рис. 1, $N^{\sigma}_{e.s.} = N^{\theta_1}_{e.s.} = 10\,000$, що при довжині ланцюга $N = 10\,000$ зайвий раз підкреслює відсутність автоколеляції, адже кожне значення містить корисну інформацію про стаціонарний розподіл.

Таким чином, наостанок наведемо деякі статистичні характеристики отриманих апостеріорних розподілів з точністю до чотирьох значущих знаків (Табл. 3), зокрема вкажемо значення «класичної» Баєсової оцінки як умовного математичного сподівання.

Parameter	Empirical mean	Standard deviation	Standard error
σ	0.5768	0.2054	0.002054
θ_1	1.3640	0.1139	0.001139

Таблиця 3: Результати імплементації вибірки Гіббса засобами `R-base`

4.3 `R-jags` Gibbs Sampler

Тепер скористаємося засобами мови `R`, пакетом `rjags`, для реалізації вибірки Гіббса. Вибірку даних (1) задамо з файлу `hw5.csv`. Покладемо $n = k = 3$ та $m = 100$, тобто матимемо Баєсову модель вигляду

$$X_{ij} \,|\, \theta_i \overset{\text{ind}}{\sim} \text{Poiss}(\theta_i), \;\; i = \overline{1,3}, \;\; j = \overline{1,100}, \tag{1}$$

$$\theta_i \,|\, \sigma \overset{\text{i.i.d.}}{\sim} \text{Gamma}(3, \sigma), \;\; i = \overline{1,3}, \tag{2}$$

$$\sigma^{-1} \sim \text{Exp}(1) \tag{3}$$

Перш ніж рухатися далі, наголосимо, що параметр σ у Гамма-розподілі є «scale parameter», в той час як засаби `rjags` надають можливість задати модель виключно через «rate parameter». У такому разі, еквівалентний запис моделі матиме вид:

$$X_{ij} \,|\, \theta_i \overset{\text{ind}}{\sim} \text{Poiss}(\theta_i), \;\; i = \overline{1,3}, \;\; j = \overline{1,100}, \tag{4}$$

$$\theta_i \,|\, \sigma \overset{\text{i.i.d.}}{\sim} \text{Gamma}(3, \sigma^{-1}), \;\; i = \overline{1,3}, \tag{5}$$

$$\sigma^{-1} \sim \text{Exp}(1) \tag{6}$$

У такому разі, запис моделі через синтаксис `rjags` наведено на Лістингу 5, а безпосередній запуск алгоритму вказано у Лістингу 6.

Лістинг 5: Задана модель засобами `rjags`

```
jags_model_string <- "model {
    for (i in 1:(n * m)) {
        x[i] ~ dpois(theta[group[i]])
    }

```

```
    for (j in 1:n) {
        theta[j] ~ dgamma(k, invsigma)
    }

    invsigma ~ dexp(1.0)
}"
```

Лістинг 6: Запуск вибірки Гіббса засобами `rjags`

```
# Convert three flows of data (x1,x2,x3) into one list and label each datapoint
                                        accordingly (group 1.0, 2.0 or 3.0)
x_jags <- array(dim = c(m * n, 2))
colnames(x_jags) <- c("x", "group")

x_jags[1:100, 1] <- x[, 1]
x_jags[1:100, 2] <- 1.0

x_jags[101:200, 1] <- x[, 2]
x_jags[101:200, 2] <- 2.0

x_jags[201:300, 1] <- x[, 3]
x_jags[201:300, 2] <- 3.0

data_jags <- list(
    x = x_jags[, 1],
    group = x_jags[, 2],
    n = n,
    m = m,
    k = prior$k
)

model <- jags.model(
    textConnection(jags_model_string),
    data = data_jags,
    n.chains = 3,
)

update(model, 100) # burn-in period

params <- c("theta", "invsigma")

model_simulation <- coda.samples(
    model = model,
    variable.names = params,
    n.iter = 10e3
)
```

В рамках додаткового дослідження збіжності MCMC було запущено три ланцюги з різних початкових точок відліку. У першому наближенні з Рис. 3 можна зробити висновок, що бажана збіжність досягнута.

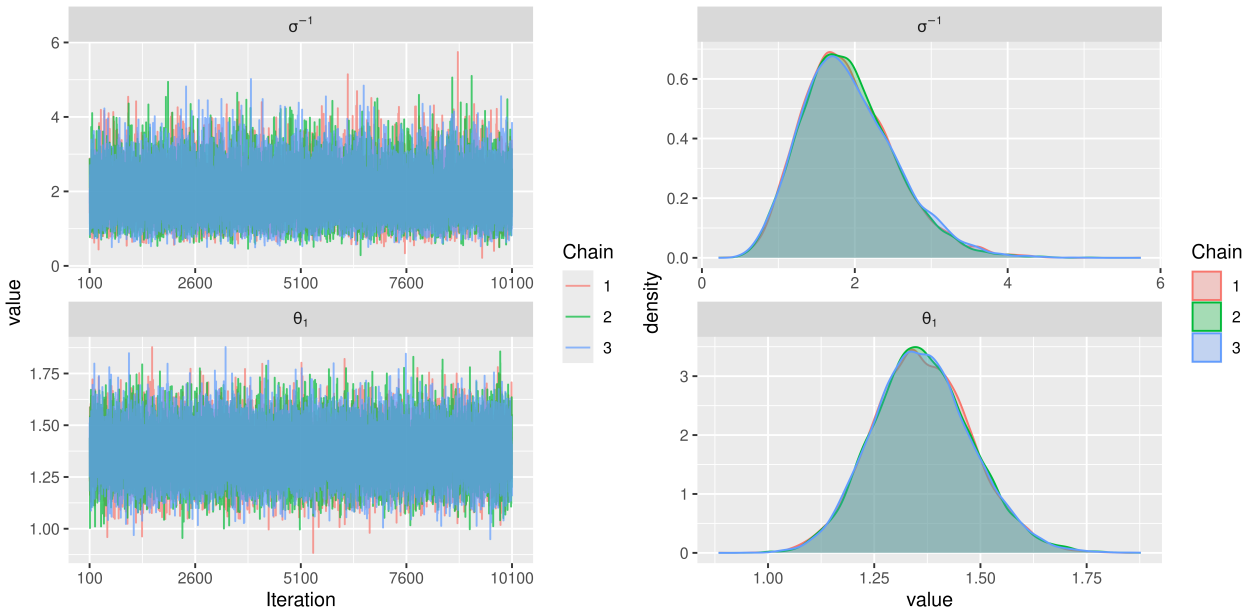

Рис. 3: Апостеріорний розподіл параметрів σ та θ_1 (засоби `R-jags`)

Тим не менш, додатковим кроком проведемо діагностику Гелмана-Рубіна (Gelman-Rubin diagnostic), яка полягає у визначенні своєрідної міри розрізненості між декількома ланцюгами [13]. Діагностика обчислює варіабельність (розкид, дисперсію) всередині ланцюгів, порівнюючи її з дисперсією між ланцюгами. Якщо всі ланцюги наблизилися до стаціонарного розподілу, то варіабельність між ланцюгами має бути відносно невеликою, а так званий «коефіцієнт потенційного зменшення масштабу» (potential scale reduction factor), який є результатом діагностики, має бути близьким до одиниці. Якщо значення значно перевищує одиницю, то можна зробити висновок, що ланцюги ще не збіглися за задану кількість ітерацій.

Рис. 4: Діагностика Гелмана-Рубіна для параметрів σ та θ_1

Бачимо (Рис. 4), що значення конфіцієнта близьке до одиниці як для параметра σ, так і для параметра θ_1. Більш того, спостерігаємо, що бажана збіжність досягнута вже на початкових ітераціях, тому «період розгону» ланцюга (burn-in period) було обачно покласти невеликим: 100 перших ітерацій (Лістинг 6).

Таким чином, у підсумку розглянемо перший з трьох ланцюгів, для якого дослідимо ознаки автокореляції (Рис. 5) та побудуємо Баєсівські оцінки (Табл. 4), які можна вважати змістовними та нідійними, адже ефективний розмір вибірки (effective size) є хоч і нижчим, ніж у попередньому підрозділі, однак все ще досить значним: $N_{e.s.}^{\sigma^{-1}} = 3628$, $N_{e.s.}^{\theta_1} = 5946$.

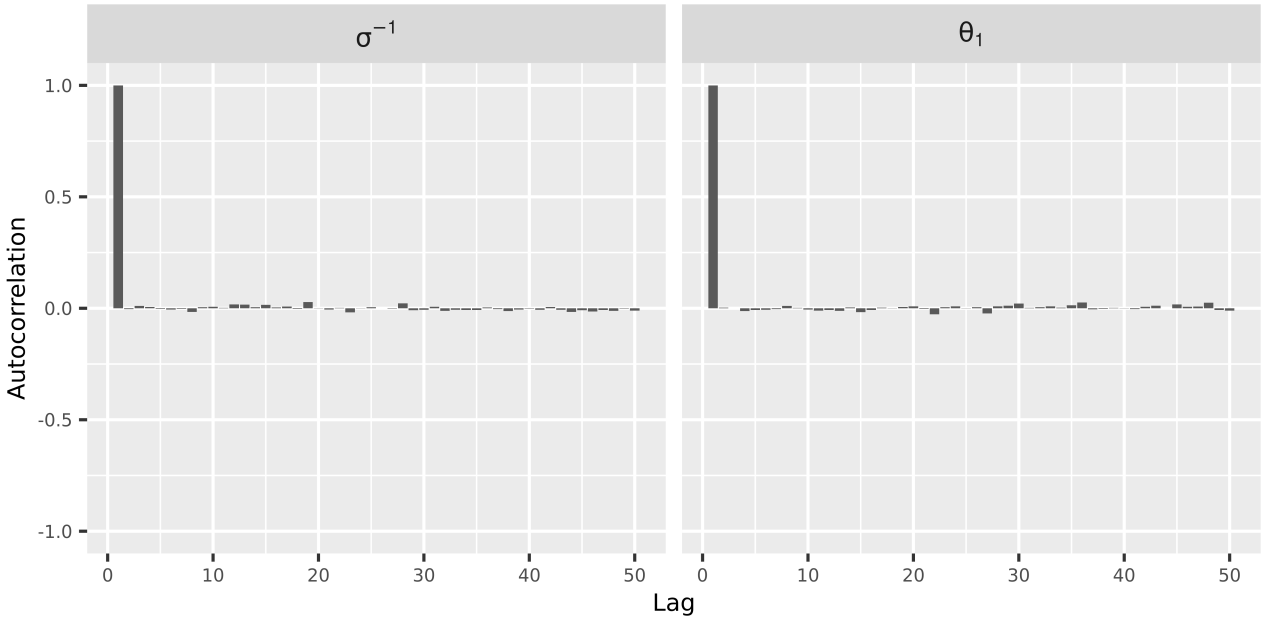

Рис. 5: Автокореляційний графік параметрів σ та θ_1 (засоби `R-jags`)

Parameter	Empirical mean	Standard deviation	Standard error
σ^{-1}	1.9320	0.6071	0.006071
θ_1	1.3660	0.1164	0.001164

Таблиця 4: Результати імплементації вибірки Гіббса засобами `R-jags`

Висновки до розділу

У цьому розділі детально було розглянуто ієрархічну Баєсову модель (як частину емпіричних моделей), її особливості використання. Показано перспективи застосування для даних, природа яких має різні джерела надходження при загальному спільному лейтмотиві (серед інтерпретацій: макроекономічні показники з різних регіонів країни, оцінки учнів з різних шкіл, кондитерська продукція з різних заводів, фізичні показники одного і того ж об'єкта на різних атомних станціях тощо).

Крім того, наведено порівняння «ручного» виведення математичної моделі та із використанням допоміжних засобів мови `R`, статистичних пакетів `rjags` та `coda` (про це детальніше буде йти мова згодом у наступному розділі). Результати, як показано, ідентичні, що дає змогу ускладнювати саму модель без зайвих кропотів щодо програмної реалізації.

5 Програмне забезпечення

Програмне забезпечення частково наведено у попередньому розділі — інструменти та засоби мови `R`, зокрема статистичні пакети `rjags` та `coda`. Щойно математичний опис моделі набуде завершального вигляду, цей розділ буде наповнено відповідними матеріалами та напрацюваннями стосовно повного опису програмного забезпечення (включаючи, можливо, опис програмних засобів статистичного пакету `Stan`).

6 Верифікація та валідація

Висновки

Перелік посилань

1. Макроекономіка: базовий курс [Електронний ресурс]: навчальний посібник / І. Й. Малий, І. Ф. Радіонова, Т. Ф. Куценко, Н. В. Федірко. — КНЕУ, 2016. — С. 254. — ISBN 978-966-926-116-8.
2. *Вебсайт.* Державна служба статистики України. — URL: https://www.ukrstat.gov.ua/.
3. *Вебсайт.* Національний банк України. — URL: https://bank.gov.ua/.
4. *Вебсайт.* Міністерство фінансів України. — URL: https://www.mof.gov.ua/uk.
5. *Вебсайт.* Інститут економіки та прогнозування НАН України. — URL: http://ief.org.ua/.
6. *Вебсайт.* Інститут демографії та соціальних досліджень ім. М. В. Птухи НАН України. — URL: https://idss.org.ua/index.
7. *Вебсайт.* Інститут регіональних досліджень ім. М. І. Долішнього НАН України. — URL: http://ird.gov.ua/.
8. *Вебсайт.* SAS. — URL: https://www.sas.com/ru_ua/home.html.
9. *Вебсайт.* IBM SPSS Statistics. — URL: https://www.ibm.com/products/spss-statistics.
10. *Вебсайт.* Stata. — URL: https://www.stata.com/.
11. *Бідюк П. І.*, *Романенко В. Д.*, *Тимощук О. Л.* Аналіз часових рядів [Електронний ресурс]: навчальний посібник. — НТУУ «КПІ»., 2010.
12. *Robert C. P.* Bayesian Model Selection and Statistical Modeling by Tomohiro Ando // International Statistical Review. — 2011. — Квіт. — Т. 79, № 1. — С. 120—121. — ISSN 0306-7734. — DOI: 10.1111/j.1751-5823.2011.00134_7.x. — URL: http://dx.doi.org/10.1111/j.1751-5823.2011.00134_7.x.
13. *McElreath R.* Statistical Rethinking: A Bayesian Course with Examples in R and Stan. — Chapman, Hall/CRC, 01.2018. — ISBN 9781315372495. — DOI: 10.1201/9781315372495. — URL: http://dx.doi.org/10.1201/9781315372495.