

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (государственный университет)
ФИЗТЕХ-ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
КАФЕДРА «ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ»

Карпев Глеб Андреевич

**Порождающие модели для прогнозирования наборов
временных рядов**

01.03.02 Прикладная математика и информатика

Выпускная квалификационная работа бакалавра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2025

Содержание

1 Введение	4
2 Основные обозначения и сокращения	4
3 Постановка задачи	5
4 Обзор литературы	6
4.1 Генеративные модели на основе оценки градиента (SGM)	6
4.2 Геометрия многообразия SPD-матриц	8
4.2.1 Аффинно-инвариантная метрика и расстояние	8
4.2.2 Касательные пространства и отображения	8
4.3 Римановы генеративные модели (RSGM)	9
5 Предложенный метод и его реализации	11
5.1 Подход 1: Условная SGM в евклидовом пространстве	11
5.1.1 Архитектура условной Score-сети	11
5.1.2 Алгоритмы обучения и прогнозирования	11
5.1.3 Ограничения евклидова подхода	12
5.2 Подход 2: Условная RSGM на многообразии SPD-матриц	12
5.2.1 Архитектура римановой Score-сети	12
5.2.2 Алгоритмы обучения и прогнозирования на многообразии	13
6 Вычислительные эксперименты	14
6.1 Описание данных и постановка эксперимента	14
6.1.1 Генерация синтетических данных	14
6.1.2 Постановка эксперимента и метрики качества	14
6.2 Сравниваемые модели	15
6.3 Результаты и их анализ	15
6.3.1 Количественное сравнение моделей	15
6.3.2 Визуальный анализ сгенерированных матриц	16
7 Заключение	18
Список литературы	19

Аннотация

Решается задача прогнозирования наборов временных рядов с высокой ковариацией и высокой дисперсией. Для решения данной задачи предлагается построить пространство парных расстояний, представляющего метрическую конфигурацию временных рядов. Прогноз осуществляется в данном пространстве, а затем результат возвращается в исходное пространство с использованием метода многомерного шкалирования. В данной работе предлагаются порождающие модели для прогнозирования наборов временных рядов в метрическом вероятностном пространстве. Новизна работы заключается в применении римановых моделей для прогнозирования и использовании римановых генеративных диффузных моделей.

Ключевые слова: *многообразие SPD-матриц, Riemannian Score-based Generative Models, conditional generation*

1 Введение

В данной работе предлагается метод прогнозирования набора временных рядов с высокой ковариацией и высокой дисперсией. Предлагаемое решение задачи прогнозирования состоит из трех этапов. Во-первых, построение пространства парных расстояний, где используемая метрика удовлетворяет условию Мерсера. Во-вторых, прогнозируется матрица попарных расстояний. В-третьих, результат возвращается в исходное пространство. В данной работе изучаются римановы генеративные диффузные модели (RSGMs) [1] для выполнения прогнозирования матрицы попарных расстояний.

Классическими алгоритмами прогнозирования временных рядов являются метод SSA (Singular Spectrum Analysis) [2], рекуррентные нейронные сети с долгой краткосрочной памятью (LSTM) [3], State space model [4]. Новизна предложенного метода заключается в том, что выполняется кодирование временных рядов с помощью матрицы расстояний, выполняется прогноз, а затем декодирование полученной матрицы.

Генеративные диффузионные модели, также называемые Score-based generative models (SGMs), представляют собой мощный класс порождающих моделей [5, 6]. SGM состоит из этапа зашумления, на котором к данным постепенно добавляется гауссовский шум, и процесса шумоподавления, на котором генеративная модель постепенно снижает гауссовский шум, чтобы получить образец данных. Существующие SGM предполагают, что данные находятся в евклидовом пространстве, но зачастую данные естественным образом описываются с помощью римановых многообразий, и методы SGM не подходят. Предлагается использовать римановы генеративные диффузные модели (RSGMs) [1], класс генеративных моделей, расширяющих SGM [5, 6] до римановых многообразий.

Анализ предлагаемого метода прогнозирования проводится на синтетических и реальных данных. Синтетический набор данных построен на основе синусоидальных сигналах со случайной амплитудой и частотой. Реальные данные представляют собой временной ряд цены на электроэнергию. Цель эксперимента — показать работоспособность предложенного подхода и провести его сравнение с базовым алгоритмом. В качестве базовой модели для сравнения используется модель, в которой прогнозное значение ковариационной матрицы на следующем шаге полагается равным ее значению на предыдущем шаге. Получена оценка качества работы построенного алгоритма.

Данная работа имеет следующую структуру. Раздел 2 вводит основные обозначения. В Разделе 3 приводится формальная постановка задачи. Раздел 4 представляет обзор по генеративным моделям и основам римановой геометрии. В Разделе 5 детально описывается предложенный метод с использованием условной генерации. Раздел 6 содержит описание и результаты вычислительных экспериментов. В Разделе 7 подводятся итоги работы и обсуждаются дальнейшие направления исследований.

2 Основные обозначения и сокращения

В тексте настоящей работы приняты следующие математические обозначения. Скалярные величины обозначаются строчными буквами курсивом, например, t или n .

Векторы представляются строчными полужирными буквами, например, \mathbf{x}_t . Матрицы обозначаются прописными полужирными буквами, например, \mathbf{C}_t . Операция транспонирования вектора или матрицы обозначается надстрочным индексом \mathbf{T} .

Помимо математических символов, в работе используется ряд сокращений для обозначения ключевых понятий и методов. Сокращение SGM (Score-based Generative Models) используется для обозначения порождающих моделей на основе оценки градиента. Их обобщение на случай неевклидовых пространств обозначается как RSGM (Riemannian Score-based Generative Models).

Ключевым объектом в предлагаемом методе являются симметричные положительно определенные матрицы, для которых используется аббревиатура SPD (Symmetric Positive-Definite).

Для обозначения методов спектрального анализа временных рядов, упоминаемых в обзоре литературы, используются сокращения SSA (Singular Spectrum Analysis) и MSSA (Multivariate Singular Spectrum Analysis) для одномерного и многомерного случаев соответственно.

3 Постановка задачи

Рассматривается задача прогнозирования набора из n временных рядов. В каждый дискретный момент времени t состояние системы представлено в виде вектора-столбца $\mathbf{x}_t = [x_{1t}, \dots, x_{nt}]^T$. Предполагается, что временные ряды в наборе характеризуются высокой степенью взаимной корреляции и значительной дисперсией, что усложняет их совместное моделирование.

Основная идея предлагаемого подхода заключается в переходе от прямого прогнозирования векторов \mathbf{x}_t к прогнозированию их совместной статистической структуры. Для этого в каждый момент времени t совокупность исторических данных о наборе временных рядов ставится в соответствие с матрицей ковариаций \mathbf{C}_t , которая вычисляется по следующей формуле:

$$\mathbf{C}_t = \frac{1}{t-1} \mathbf{Y}_t \mathbf{Y}_t^T, \quad (1)$$

где \mathbf{Y}_t — это матрица, столбцы которой представляют собой центрированные векторы временных рядов:

$$\mathbf{Y}_t = [\mathbf{x}_1 - \mu_t, \dots, \mathbf{x}_t - \mu_t]. \quad (2)$$

Здесь μ_t является вектором среднего значения, вычисленным по всем доступным историческим данным до момента t :

$$\mu_t = \frac{1}{t} \sum_{i=1}^t \mathbf{x}_i. \quad (3)$$

Таким образом, задача сводится к нахождению оптимального отображения f , действующего в пространстве симметричных положительно определенных матриц размерности $n \times n$, которое обозначается как $Sym^+(n)$. Это отображение

$$f : Sym^+(n) \rightarrow Sym^+(n)$$

должно прогнозировать будущую ковариационную матрицу $\mathbf{C}_{t+\Delta}$ на основе текущей \mathbf{C}_t . Качество прогноза $\hat{\mathbf{C}}_{t+\Delta} = f(\mathbf{C}_t)$ оценивается с помощью метрики, учитывающей

внутреннюю геометрию пространства $Sym^+(n)$. В качестве функции потерь используется ожидание квадрата геодезического расстояния между истинной и спрогнозированной матрицами. Задача оптимизации формулируется как поиск такого отображения f , которое минимизирует следующую функцию ошибки:

$$\mathcal{L} = \mathbb{E} \left[d_{\text{Riem}}^2 \left(\mathbf{C}_{t+\Delta}^{\text{true}}, f(\mathbf{C}_t) \right) \right] \rightarrow \min_f, \quad (4)$$

где $d_{\text{Riem}}(\cdot, \cdot)$ обозначает риманово геодезическое расстояние на многообразии $Sym^+(n)$.

4 Обзор литературы

В данной главе рассматриваются теоретические основы, составляющие фундамент предложенного в работе метода. Сначала будет дан обзор генеративных моделей на основе оценки градиента (SGM), затем будут введены ключевые понятия из римановой геометрии, касающиеся многообразия симметричных положительно определенных (SPD) матриц, и, наконец, будет описано их обобщение — римановы генеративные модели (RSGM).

4.1 Генеративные модели на основе оценки градиента (SGM)

Генеративные модели на основе оценки градиента (Score-based Generative Models, SGM), также известные как диффузионные модели, представляют собой мощный класс порождающих моделей, продемонстрировавших передовые результаты в различных областях [5, 7]. Методология SGM состоит из двух ключевых процессов, которые описываются стохастическими дифференциальными уравнениями (SDE) [8].

Прямой процесс (зашумление). Этот процесс постепенно преобразует сложное распределение данных $p_0(\mathbf{x})$ в простое, априорное распределение $p_T(\mathbf{x})$, как правило, в стандартное гауссовское $\mathcal{N}(0, \mathbf{I})$. Процесс описывается SDE общего вида:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{W}, \quad (5)$$

где $t \in [0, T]$, $f(\mathbf{x}, t)$ — векторная функция дрейфа (drift), $g(t)$ — скалярная функция диффузии (diffusion), а \mathbf{W} — стандартный винеровский процесс. Функции f и g выбираются таким образом, чтобы при $t \rightarrow T$ распределение $p_t(\mathbf{x})$ сходилось к априорному, а переходное распределение $p_{t|s}(\mathbf{x}_t | \mathbf{x}_s)$ для $s < t$ было гауссовским и имело аналитическое выражение.

Обратный процесс (генерация). Согласно теореме Андерсона [9], существует обратное по времени SDE, которое описывает процесс, обращающий диффузию, заданную уравнением (5). Это обратное SDE имеет вид:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt + g(t)d\bar{\mathbf{W}}, \quad (6)$$

где $\bar{\mathbf{W}}$ — винеровский процесс при обратном течении времени, а $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ — это градиент логарифма плотности данных в момент времени t . Этот градиент называется *функцией оценки* (score function).

Решение этого обратного уравнения, начиная с сэмпла $\mathbf{x}_T \sim p_T$, позволяет сгенерировать новый объект из исходного распределения данных $\mathbf{x}_0 \sim p_0$.

Аппроксимация функции оценки. Ключевая сложность заключается в том, что функция оценки $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ неизвестна, так как распределение $p_t(\mathbf{x})$ недоступно в явном виде. Для ее аппроксимации используется нейронная сеть $s_{\theta}(\mathbf{x}, t)$, обучаемая с помощью метода, известного как Denoising Score Matching [10]. Целевая функция для обучения нейросети имеет вид:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t} \left[\lambda(t) \left\| s_{\theta}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t | \mathbf{x}_0) \right\|^2 \right], \quad (7)$$

где математическое ожидание берется по времени $t \in [0, T]$, сэмплам из исходного распределения $\mathbf{x}_0 \sim p_0(\mathbf{x})$ и сэмплам $\mathbf{x}_t \sim p_{t|0}(\mathbf{x}_t | \mathbf{x}_0)$. Преимущество этого подхода заключается в том, что условная плотность $p_{t|0}(\mathbf{x}_t | \mathbf{x}_0)$ известна аналитически (является гауссовской), и ее градиент легко вычисляется. Функция $\lambda(t)$ является весовой функцией, которая влияет на качество генерации.

Основным ограничением стандартных SGM является то, что все операции, включая вычисление градиента, определены в евклидовом пространстве \mathbb{R}^n . Это делает их неприменимыми для данных, которые естественным образом существуют на многообразиях с нетривиальной геометрией. Общая схема работы SGM представлена на рисунке 1.

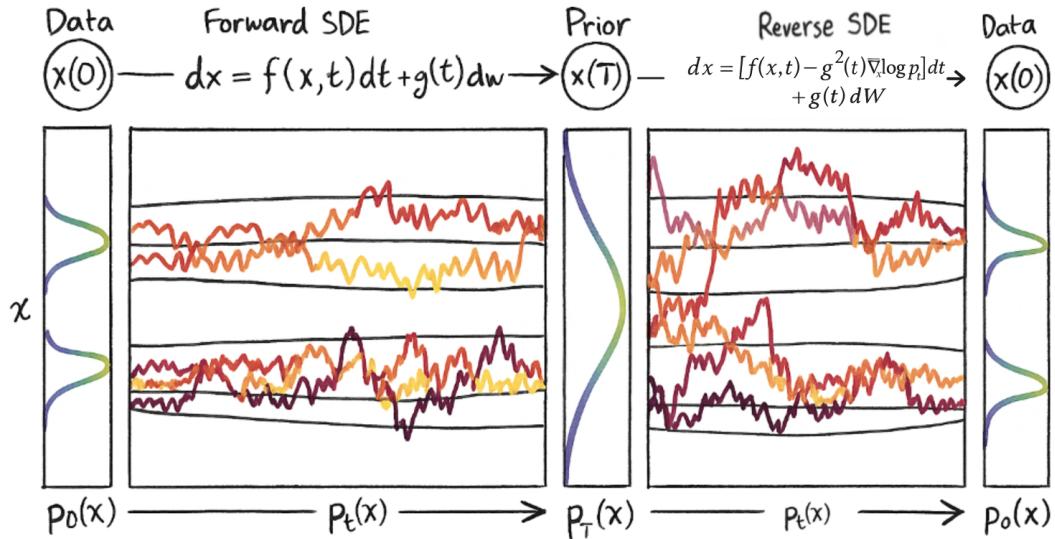


Рис. 1: Схематическое представление работы Score-based Generative Model (SGM). Прямой процесс постепенно добавляет шум к исходным данным $p_0(\mathbf{x})$, превращая их в простое гауссовское распределение $p_T(\mathbf{x})$. Обратный процесс начинается со случайног шума и, следуя оценкам градиента $\nabla \log p_t$, выученным моделью s_{θ} , восстанавливает объект из исходного распределения данных.

4.2 Геометрия многообразия SPD-матриц

Ковариационные матрицы, являющиеся центральным объектом в данной работе, по определению являются симметричными и положительно определенными (Symmetric Positive-Definite, SPD). Множество всех SPD-матриц размера $n \times n$ образует риманово многообразие. Далее в тексте мы будем обозначать это многообразие как $\mathcal{M} = Sym^+(n)$ или, следуя обозначениям на рисунке 2, как \mathcal{S}_{++} . Это многообразие обладает богатой геометрической структурой, которую необходимо учитывать для корректного анализа и моделирования данных [11, 12]. Ключевое отличие заключается в том, что кратчайшим путем между двумя точками (матрицами) является кривая, называемая геодезической.

4.2.1 Аффинно-инвариантная метрика и расстояние

Центральным элементом риманова многообразия является метрический тензор, который определяет способ измерения расстояний и углов. Для многообразия \mathcal{M} стандартным выбором является аффинно-инвариантная метрика. Она обладает важным свойством инвариантности относительно аффинных преобразований исходных данных, что делает ее особенно подходящей для ковариационных матриц. Скалярное произведение двух векторов \mathbf{A}, \mathbf{B} , принадлежащих касательному пространству в точке $\mathbf{G} \in \mathcal{M}$, определяется как:

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{G}} = \text{tr}(\mathbf{G}^{-1} \mathbf{A} \mathbf{G}^{-1} \mathbf{B}). \quad (8)$$

На основе этой метрики определяется геодезическое расстояние, обозначаемое на рисунке 2 как $\delta(\cdot, \cdot)$. Расстояние между матрицами \mathbf{C}_1 и \mathbf{G} вычисляется по формуле:

$$\delta_{\text{Riem}}(\mathbf{C}_1, \mathbf{G}) = \|\text{Log}_{\mathbf{G}}(\mathbf{C}_1)\|_F = \|\log(\mathbf{G}^{-1/2} \mathbf{C}_1 \mathbf{G}^{-1/2})\|_F, \quad (9)$$

где $\text{Log}_{\mathbf{G}}(\mathbf{C}_1)$ — логарифмическое отображение, $\log(\cdot)$ — операция взятия главного матричного логарифма, а $\|\cdot\|_F$ — норма Фробениуса.

4.2.2 Касательные пространства и отображения

Для выполнения вычислений на многообразии необходимо уметь переходить от его криволинейной структуры к локальным приближениям. Эти пространства называются касательными пространствами. Касательное пространство $T_{\mathbf{G}}\mathcal{M}$ в точке \mathbf{G} можно интуитивно представить как плоскость, касающуюся многообразия в этой точке. Векторы в этом пространстве представляют собой симметричные матрицы, обозначаемые на рисунке 2 как \mathcal{S} . Переход между многообразием (\mathcal{S}_{++}) и его касательными пространствами (\mathcal{S}) осуществляется с помощью двух ключевых взаимно обратных операций, которые схематически показаны на рисунке 2.

Логарифмическое отображение (Log). Данное отображение переносит точку \mathbf{C} с многообразия в касательное пространство в точке \mathbf{G} , представляя ее в виде вектора ξ , который указывает направление и «длину» геодезического пути от \mathbf{G} до \mathbf{C} . Формула логарифмического отображения имеет вид:

$$\xi = \text{Log}_{\mathbf{G}}(\mathbf{C}) = \mathbf{G}^{1/2} \log(\mathbf{G}^{-1/2} \mathbf{C} \mathbf{G}^{-1/2}) \mathbf{G}^{1/2}. \quad (10)$$

Экспоненциальное отображение (Exp). Данное отображение, обратное логарифмическому, «перемещает» точку \mathbf{G} вдоль геодезической в направлении, заданном вектором ξ из касательного пространства $T_{\mathbf{G}}\mathcal{M}$. Результатом является новая точка на многообразии. Формула экспоненциального отображения:

$$\mathbf{C} = \text{Exp}_{\mathbf{G}}(\xi) = \mathbf{G}^{1/2} \exp \left(\mathbf{G}^{-1/2} \xi \mathbf{G}^{-1/2} \right) \mathbf{G}^{1/2}, \quad (11)$$

где $\exp(\cdot)$ — операция взятия матричной экспоненты. Эти два отображения являются основой для всех вычислений в предлагаемом методе, позволяя корректно работать с геометрией пространства ковариационных матриц.

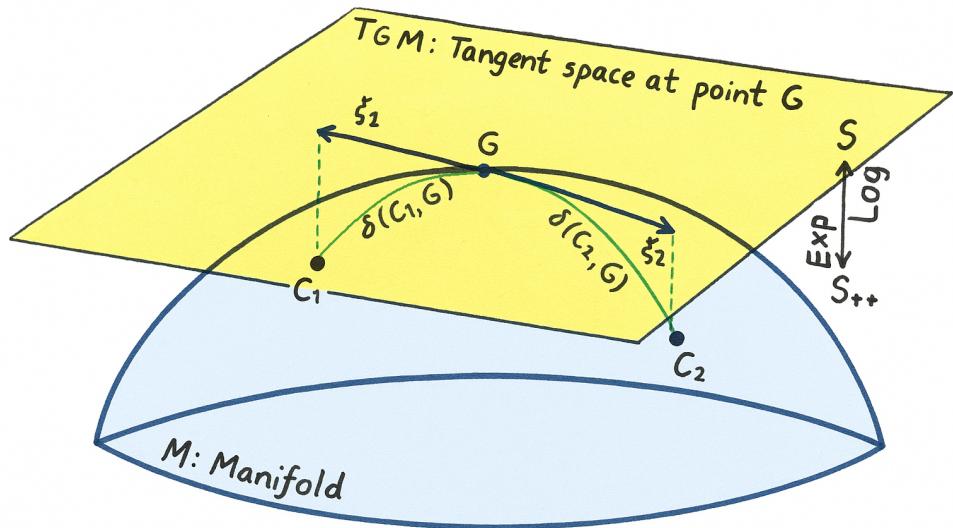


Рис. 2: Концептуальная иллюстрация риманова многообразия \mathcal{M} (также \mathcal{S}_{++} , синяя поверхность) и его касательного пространства $T_{\mathbf{G}}\mathcal{M}$ (также \mathcal{S}) в точке \mathbf{G} . Логарифмическое отображение (Log) переносит точки с многообразия (например, C_1) в векторы касательного пространства (например, ξ_1). Экспоненциальное отображение (Exp) выполняет обратную операцию. Зеленые линии $\delta(C_1, G)$ и $\delta(C_2, G)$ — геодезические пути на многообразии.

4.3 Римановы генеративные модели (RSGM)

Стандартные генеративные модели на основе оценки градиента (SGM), как было показано ранее, по своей конструкции ограничены евклидовыми пространствами. Для преодоления этого ограничения и работы с данными на многообразиях были разработаны римановы генеративные модели (RSGM) [1]. Идея RSGM заключается в том, чтобы определить прямой и обратный диффузионные процессы непосредственно на римановом многообразии \mathcal{M} , используя его внутреннюю геометрию, описанную в разделе 4.2.

Прямой процесс на многообразии. Аналогично евклидову случаю, прямой процесс в RSGM постепенно разрушает преобразует исходное распределение $p_0(\mathbf{C})$

на многообразии в априорное распределение. Этот процесс описывается стохастическим дифференциальным уравнением (SDE), определенным непосредственно на многообразии \mathcal{M} :

$$d\mathbf{C}_t = b(\mathbf{C}_t)dt + \sigma(t)d\mathbf{B}_t^{\mathcal{M}}, \quad (12)$$

где $b(\cdot)$ — векторное поле дрейфа на многообразии, $\sigma(t)$ — скалярная функция диффузии, а $d\mathbf{B}_t^{\mathcal{M}}$ — приращение риманова броуновского движения. Риманово броуновское движение представляет собой случайное блуждание, траектория которого ограничена поверхностью многообразия, что схематично показано на рисунке 3.

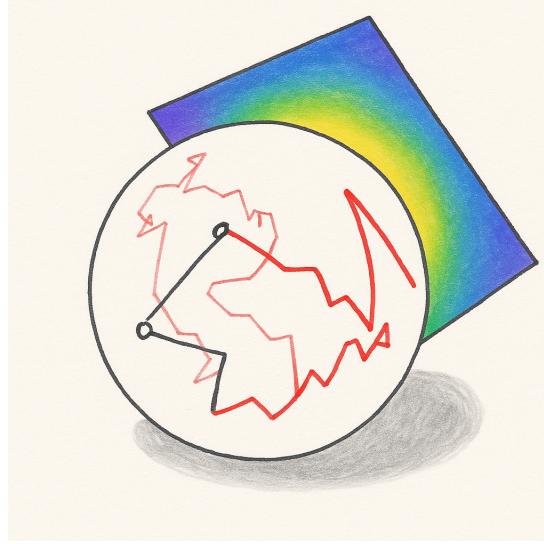


Рис. 3: Концептуальная иллюстрация траектории риманова броуновского движения на многообразии (сфере). В отличие от стандартного броуновского движения, случайное блуждание происходит по поверхности многообразия, следуя его кривизне.

Обратный процесс на многообразии. Как и в евклидовом случае, существует обратное по времени SDE, которое позволяет восстанавливать данные из шума. Уравнение для обратного процесса на многообразии \mathcal{M} имеет следующий вид [1, Theorem 3.1]:

$$d\mathbf{C}_t = [-b(\mathbf{C}_t) + \sigma(t)^2 \nabla_{\mathcal{M}} \log p_t(\mathbf{C}_t)] dt + \sigma(t)d\bar{\mathbf{B}}_t^{\mathcal{M}}, \quad (13)$$

где $\nabla_{\mathcal{M}} \log p_t(\mathbf{C}_t)$ — это *риманова функция оценки* (Riemannian score function), то есть градиент логарифма плотности, вычисленный на многообразии. Это ключевое отличие от стандартных SGM, где используется евклидов градиент.

Аппроксимация римановой функции оценки. Основная задача в RSGM, как и в SGM, заключается в аппроксимации неизвестной функции оценки. Для этого используется нейронная сеть $s_{\theta}(\mathbf{C}, t)$, которая обучается предсказывать риманов градиент. Поскольку операции на многообразии нелинейны, нейронная сеть, как правило, работает в касательных пространствах. Это означает, что модель s_{θ} принимает на вход точку на многообразии \mathbf{C} и предсказывает вектор в касательном пространстве $T_{\mathbf{C}}\mathcal{M}$, который аппроксимирует истинную функцию оценки.

Такой подход гарантирует, что весь процесс генерации происходит на многообразии, а сгенерированные объекты (в нашем случае ковариационные матрицы) по

построению сохраняют свои структурные свойства, такие как симметричность и положительная определенность.

В следующей главе на основе этих концепций будет подробно описан предложенный метод прогнозирования наборов временных рядов, реализующий условную генерацию ковариационных матриц с помощью RSGM.

5 Предложенный метод и его реализации

В данной главе подробно описываются два подхода к реализации условной генеративной модели для прогнозирования ковариационных матриц. Первый подход основан на стандартной евклидовой модели SGM, где матрица рассматривается как обычный тензор в евклидовом пространстве. Второй, основной, подход использует римановы SGM (RSGM), которые явным образом учитывают геометрию пространства SPD-матриц. Цель состоит в сравнении этих двух парадигм для решения поставленной задачи прогнозирования условного распределения $p(\mathbf{C}_{t+\Delta}|\mathbf{C}_t)$.

5.1 Подход 1: Условная SGM в евклидовом пространстве

Первая реализация представляет собой адаптацию стандартной диффузионной модели, в которой ковариационные матрицы рассматриваются как двумерные изображения. Этот подход игнорирует внутреннюю геометрию многообразия $Sym^+(n)$ и работает исключительно в евклидовом пространстве $\mathbb{R}^{n \times n}$. Данная модель была реализована на фреймворке PyTorch.

5.1.1 Архитектура условной Score-сети

В качестве аппроксиматора функции оценки s_θ используется сверточная нейронная сеть архитектуры U-Net [13]. Такая архитектура хорошо зарекомендовала себя в задачах генерации изображений и подходит для работы с матрицами как с тензорами.

- **Условный вход:** Для реализации условной генерации на вход модели подается тензор, полученный конкатенацией по канальному измерению матрицы условия \mathbf{C}_t и запущенной целевой матрицы $\mathbf{C}_{t+\Delta}(\tau)$. Таким образом, входной тензор имеет размерность $(B, 2, n, n)$, где B — размер батча, а 2 — количество входных каналов.
- **Симметризация выхода:** Поскольку ковариационные матрицы симметричны, для того чтобы выход сети также был симметричной матрицей, к последнему слою применяется операция симметризации: $\mathbf{S}_{\text{out}} = (\mathbf{S} + \mathbf{S}^T)/2$.

5.1.2 Алгоритмы обучения и прогнозирования

Процессы обучения и прогнозирования для евклидовой SGM формализованы в алгоритмах 1 и 2.

Algorithm 1 Обучение условной SGM

Require: Обучающая выборка пар $(\mathbf{C}_t, \mathbf{C}_{t+\Delta})$.

- 1: **repeat**
 - 2: Взять случайную пару $(\mathbf{C}_t, \mathbf{C}_{t+\Delta})$ из выборки.
 - 3: Выбрать случайное время диффузии $\tau \sim \mathcal{U}(0, T)$.
 - 4: Сгенерировать случайный шум $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.
 - 5: Создать зашумленную целевую матрицу $\mathbf{C}_{t+\Delta}(\tau) = \alpha_\tau \mathbf{C}_{t+\Delta} + \sigma_\tau \mathbf{z}$.
 - 6: Сформировать входной тензор $\mathbf{X}_{\text{in}} = \text{concat}(\mathbf{C}_t, \mathbf{C}_{t+\Delta}(\tau))$.
 - 7: Вычислить предсказание модели $s_\theta(\mathbf{X}_{\text{in}}, \tau)$.
 - 8: Вычислить ошибку как $\mathcal{L} = \|s_\theta(\cdot) + \mathbf{z}/\sigma_\tau\|^2$.
 - 9: Выполнить шаг градиентного спуска для обновления параметров θ .
 - 10: **until** условие остановки выполнено
-

Algorithm 2 Прогнозирование с помощью условной SGM

Require: Обученная модель s_θ , матрица-условие \mathbf{C}_t .

- 1: Взять начальный семпл из априорного распределения: $\hat{\mathbf{C}}_T \sim \mathcal{N}(0, \mathbf{I})$.
 - 2: Решить ОДУ на интервале $[T, \varepsilon]$ с начальным условием $\hat{\mathbf{C}}_T$, используя $s_\theta(\cdot, \cdot, \mathbf{C}_t)$.
 - 3: Получить решение $\hat{\mathbf{C}}_0$.
 - 4: Спроектировать $\hat{\mathbf{C}}_0$ на пространство SPD-матриц: $\hat{\mathbf{C}}_{t+\Delta} = \text{ProjectToPSD}(\hat{\mathbf{C}}_0)$.
 - 5: **return** $\hat{\mathbf{C}}_{t+\Delta}$.
-

5.1.3 Ограничения евклидова подхода

Важной особенностью и фундаментальным ограничением данного подхода является то, что положительная определенность сгенерированных матриц не гарантируется. Модель, работая в $\mathbb{R}^{n \times n}$, не имеет информации о структурных свойствах пространства SPD-матриц. Как показано в Алгоритме 2 (шаг 4), для решения этой проблемы к сгенерированному результату применяется проекция путем обнуления отрицательных собственных значений в его спектральном разложении. Этот шаг является внешней пост-обработкой и не является частью самого процесса генерации, что может приводить к искажению результатов и потере информации.

5.2 Подход 2: Условная RSGM на многообразии SPD-матриц

Второй, основной, подход основан на фреймворке римановых генеративных моделей (RSGM), который определяет все диффузионные процессы непосредственно на многообразии $\mathcal{M} = \text{Sym}^+(n)$. Данная реализация выполнена с использованием JAX, Haiku и библиотеки дифференциальной геометрии ‘geomstats’, что позволяет корректно работать с неевклидовой структурой данных.

5.2.1 Архитектура римановой Score-сети

Как и в евклидовом случае, для аппроксимации функции оценки используется архитектура U-Net, но с ключевыми отличиями, необходимыми для работы на многообразии:

- **Работа в касательном пространстве:** Все операции свертки в сети происхо-

дят над тензорами, однако финальный выход модели — предсказание score-функции — должен быть вектором в касательном пространстве $T_{\mathbf{C}_\tau} \mathcal{M}$. Поскольку касательное пространство к многообразию SPD-матриц состоит из симметричных матриц, на выходе сети принудительно устанавливается симметрия: $\mathbf{S}_{\text{out}} = (\mathbf{S} + \mathbf{S}^T)/2$.

- **Интеграция с ‘geomstats’:** Ключевым элементом реализации является то, что объект стохастического дифференциального уравнения (в коде ‘VPSDE’ или ‘Langevin’) инициализируется объектом ‘manifold’ из ‘geomstats’. Это означает, что все шаги диффузии и генерации, включая риманово броуновское движение и вычисление функции потерь, выполняются с учетом аффинно-инвариантной метрики многообразия.

5.2.2 Алгоритмы обучения и прогнозирования на многообразии

Алгоритмы обучения и прогнозирования для римановой модели принципиально отличаются от евклидова случая, так как все операции происходят на многообразии.

Algorithm 3 Обучение условной RSGM

Require: Обучающая выборка пар $(\mathbf{C}_t, \mathbf{C}_{t+\Delta})$.

- 1: **repeat**
 - 2: Взять случайную пару $(\mathbf{C}_t, \mathbf{C}_{t+\Delta})$ из выборки.
 - 3: Выбрать случайное время диффузии $\tau \sim \mathcal{U}(0, T)$.
 - 4: Применить риманов прямой процесс для получения $\mathbf{C}_{t+\Delta}(\tau)$ из $\mathbf{C}_{t+\Delta}$.
 - 5: Вычислить истинный риманов score-вектор в касательном пространстве $T_{\mathbf{C}_{t+\Delta}(\tau)} \mathcal{M}$.
 - 6: Вычислить предсказание модели $s_\theta(\mathbf{C}_{t+\Delta}(\tau), \tau, \mathbf{C}_t)$.
 - 7: Вычислить ошибку как риманову норму разности векторов в касательном пространстве.
 - 8: Выполнить шаг градиентного спуска для обновления параметров θ .
 - 9: **until** условие остановки выполнено
-

Algorithm 4 Прогнозирование с помощью условной RSGM

Require: Обученная модель s_θ , матрица-условие \mathbf{C}_t .

- 1: Взять начальный семпл $\hat{\mathbf{C}}_T$ из априорного распределения на многообразии.
 - 2: Решить риманово обратное SDE с помощью семплера (например, предиктор Эйлера-Маруямы), используя на каждом шаге $s_\theta(\cdot, \cdot, \mathbf{C}_t)$.
 - 3: Получить итоговый семпл $\hat{\mathbf{C}}_0$, который по построению является SPD-матрицей.
 - 4: **return** $\hat{\mathbf{C}}_{t+\Delta} = \hat{\mathbf{C}}_0$.
-

Главное преимущество риманова подхода заключается в том, что процесс генерации (Алгоритм 4) гарантирует, что результат $\hat{\mathbf{C}}_0$ всегда будет оставаться на многообразии $Sym^+(n)$. Это избавляет от необходимости пост-обработки в виде проекции и позволяет сохранить целостность сгенерированных данных.

6 Вычислительные эксперименты

Для проверки работоспособности и сравнения эффективности предложенных подходов (евклидова SGM и римановой RSGM) был проведен ряд вычислительных экспериментов. В данной главе описывается постановка экспериментов, используемые наборы данных, метрики качества и полученные результаты.

6.1 Описание данных и постановка эксперимента

Экспериментальное исследование проводилось на синтетических данных, что позволяет полностью контролировать их статистические свойства и проводить более точный анализ.

6.1.1 Генерация синтетических данных

Синтетический набор данных представляет собой набор из нескольких взаимосвязанных временных рядов, сгенерированных на основе синусоидальных сигналов с добавлением гауссовского шума. Для генерации использовалась функция, которая создает временной ряд по формуле:

$$x(t) = A \cdot \sin \left(\frac{2\pi k}{N} t + \varphi_0 \right) + \mathcal{N}(0, \sigma^2),$$

где A — амплитуда сигнала, k — количество периодов на всю длину ряда, N — общее число точек, t — временной шаг, φ_0 — начальная фаза, а $\sigma = A/10$ — стандартное отклонение шума.

В рамках эксперимента было сгенерировано $n = 3$ временных ряда, каждый длиной 10000 временных шагов. Для придания рядам корреляционной структуры для каждого из них использовался свой сдвиг фазы φ_0 . Далее, для этого набора временных рядов была построена последовательность ковариационных матриц \mathbf{C}_t в соответствии с методикой, описанной в разделе 3.

6.1.2 Постановка эксперимента и метрики качества

Эксперимент заключается в решении задачи условного прогнозирования "на один шаг вперед". Модели обучались на парах матриц $(\mathbf{C}_t, \mathbf{C}_{t+1})$. Полученная последовательность матриц была разделена на обучающую и тестовую выборки в соотношении 80% на 20%. На тестовой выборке для каждой входной матрицы \mathbf{C}_t генерировался прогноз $\hat{\mathbf{C}}_{t+1}$, который затем сравнивался с истинной матрицей \mathbf{C}_{t+1} .

Для оценивания качества прогнозирования использовалось среднее геодезическое расстояние на многообразии $Sym^+(n)$, которое является основной метрикой, так как оно напрямую отражает ошибку в истинном пространстве данных:

$$\text{Ошибка} = \frac{1}{M} \sum_{i=1}^M d_{\text{Riem}} \left(\mathbf{C}_{\text{true}}^{(i)}, \hat{\mathbf{C}}_{\text{pred}}^{(i)} \right), \quad (14)$$

где M — размер тестовой выборки.

6.2 Сравниваемые модели

В эксперименте участвовали три модели:

1. **Наивная модель (Baseline):** Персистентная модель, в которой прогноз на следующем шаге полагается равным текущему значению: $\hat{\mathbf{C}}_{t+1} = \mathbf{C}_t$.
2. **Условная SGM:** Модель, описанная в разделе 5.1, работающая в евклидовом пространстве.
3. **Условная RSGM:** Основная предложенная модель, описанная в разделе 5.2, работающая на многообразии SPD-матриц.

6.3 Результаты и их анализ

В данном разделе представлены результаты вычислительных экспериментов по прогнозированию ковариационных матриц на синтетических данных. Проводится количественное и качественное сравнение трех моделей: наивной (Baseline), евклидовой условной SGM и римановой условной RSGM.

6.3.1 Качественное сравнение моделей

Основным критерием оценки качества моделей служило среднее риманово геодезическое расстояние между спрогнозированными и истинными матрицами на тестовой выборке. Данная метрика наиболее адекватно отражает ошибку, поскольку вычисляется в собственном пространстве данных. Результаты сравнения представлены в таблице 1.

Таблица 1: Сравнение производительности моделей по метрике среднего риманова расстояния. Меньшее значение соответствует лучшему качеству прогноза.

Модель	Среднее риманово расстояние $\mathbb{E}[d_{\text{Riem}}]$
Наивная модель ($\hat{\mathbf{C}}_{t+1} = \mathbf{C}_t$)	1.191
Условная SGM (в $\mathbb{R}^{n \times n}$)	1.376
Условная RSGM	1.083

Анализ полученных количественных результатов позволяет сделать несколько ключевых выводов.

Превосходство риманова подхода. Основной результат эксперимента заключается в том, что предложенная модель RSGM демонстрирует наименьшую ошибку прогнозирования, достигая значения 1.083. Это значение примерно на 9% ниже, чем у сильной наивной модели (1.191), и на 21% ниже, чем у евклидовой SGM (1.376). Это убедительно доказывает основную гипотезу данной работы: явный учет геометрии пространства SPD-матриц позволяет строить более точные и адекватные прогностические модели. Модель RSGM, работая непосредственно на многообразии, способна улавливать нелинейные зависимости в динамике ковариационных структур, которые упускаются другими подходами.

Крайне важным и показательным является тот факт, что стандартная модель SGM, работающая в евклидовом пространстве, показала результат хуже не только

RSGM, но и наивной модели. Это наблюдение имеет фундаментальное значение. Оно демонстрирует, что прямое применение мощных генеративных моделей к матрицам как к обычным тензорам без учета их внутренней структуры не просто неэффективно, но и контрпродуктивно. Евклидово расстояние между матрицами является плохой аппроксимацией «расстояния» между соответствующими им статистическими распределениями. Принудительная проекция на пространство SPD-матриц на последнем шаге является пост-обработкой, которая не может компенсировать ошибки, накопленные в процессе генерации в евклидовом пространстве. Таким образом, результат SGM является не недостатком диффузионных моделей как таковых, а сильным аргументом в пользу необходимости использования геометрии-ориентированных подходов.

Эффективность наивной модели. Высокий результат наивной модели (1.191) объясняется природой используемых временных рядов, которые обладают высокой автокорреляцией. Ковариационная структура такой системы меняется плавно и не претерпевает резких скачков от одного временного шага к другому. Это делает персистентную модель сильным и труднопреодолимым базовым уровнем. Тот факт, что RSGM смогла превзойти этот уровень, говорит о ее способности улавливать тонкие, но значимые изменения в динамике системы.

6.3.2 Визуальный анализ сгенерированных матриц

Для качественной оценки результатов были визуализированы выборки из 64 ковариационных матриц: истинные матрицы из тестовой выборки и соответствующие им прогнозы, сгенерированные каждой из моделей. Такой анализ позволяет наглядно оценить, насколько хорошо модели способны воспроизвести сложную структуру данных.

На рисунке 4 показаны матрицы, сгенерированные евклидовой моделью SGM. Видно, что результаты содержат значительное количество шума и артефактов. Структура многих матриц размыта, диагональные элементы не всегда доминируют, а характерные паттерны, присущие истинным данным, часто теряются. Это визуально подтверждает высокий качественный показатель ошибки для данной модели и иллюстрирует ее неспособность выучить корректную структуру данных при работе в евклидовом пространстве.

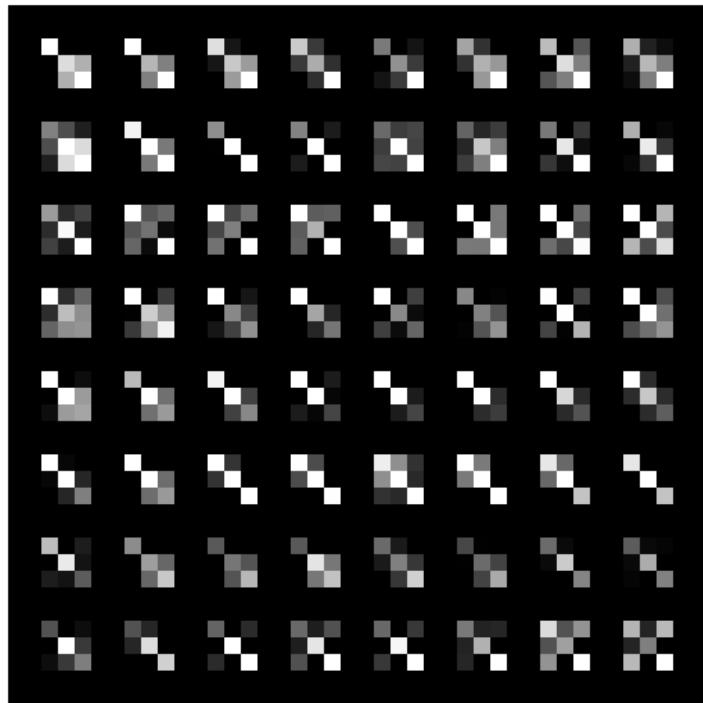


Рис. 4: Матрицы ковариаций, сгенерированные условной моделью SGM. Изображение представляет собой сетку из 64 прогнозов для соответствующей выборки из тестовых данных.

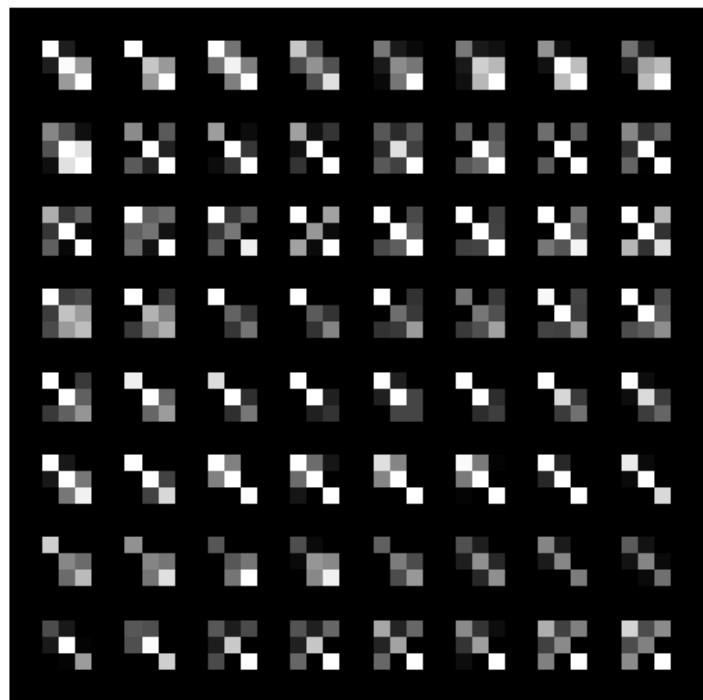


Рис. 5: Матрицы ковариаций, сгенерированные предложенной условной моделью RSGM.

Напротив, матрицы, сгенерированные римановой моделью RSGM (рисунок 5), демонстрируют высокое структурное сходство с истинными матрицами (рисунок 6). Модель RSGM успешно воспроизводит ключевые детали: доминирующую диагональ, соответствующую дисперсиям, и характерное распределение значений на побочных диагоналях, отражающее ковариационную структуру. Различия между отдельными матрицами в сетке также выглядят естественно, что говорит о способности модели улавливать тонкую динамику изменений.

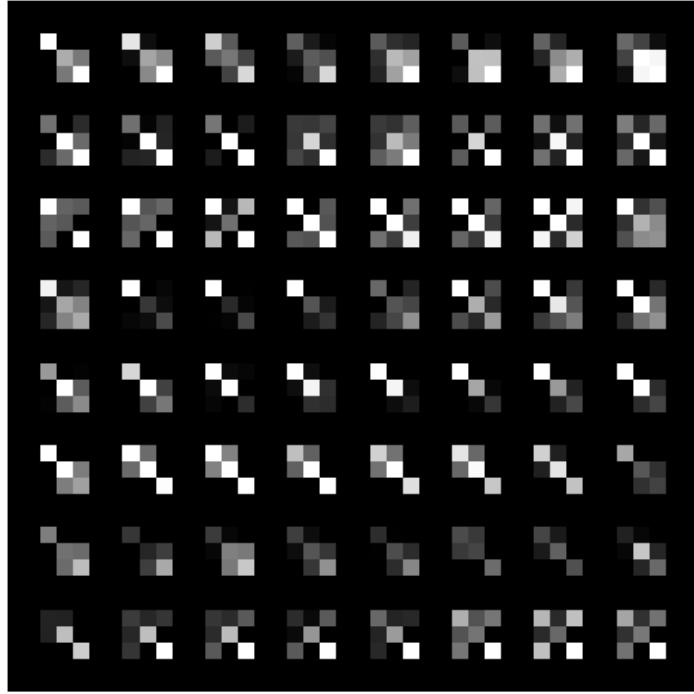


Рис. 6: Истинные ковариационные матрицы из тестовой выборки.

Наконец, на рисунке 6 представлены истинные матрицы из тестовой выборки, которые служили целью для прогнозирования. Сравнение сгенерированных изображений с этим эталоном наглядно показывает, что риманов подход позволяет получить гораздо более правдоподобные и структурно корректные результаты.

7 Заключение

В настоящей работе была исследована и решена задача прогнозирования наборов временных рядов, характеризующихся высокой ковариацией и дисперсией. Было показано, что стандартные подходы, работающие в евклидовом пространстве, могут быть неэффективны для таких данных. Для решения этой проблемы был предложен метод, основанный на представлении временных рядов в виде ковариационных матриц и последующем прогнозировании в пространстве симметричных положительно определенных (SPD) матриц с использованием его внутренней римановой геометрии.

В рамках работы были реализованы и сравнены два подхода:

- Модель, основанная на стандартной SGM, которая рассматривает матрицы как тензоры в евклидовом пространстве.

- Основная предложенная модель, основанная на условной римановой SGM (RSGM), которая определяет диффузионные процессы непосредственно на многообразии SPD-матриц.

Вычислительные эксперименты на синтетических данных убедительно продемонстрировали преимущество геометрия-ориентированного подхода. Предложенная модель RSGM показала наилучшее качество прогнозирования, достигнув среднего риманова расстояния 1.083, что существенно ниже, чем у наивной модели (1.191) и евклидовой SGM (1.376). Визуальный анализ также подтвердил, что RSGM способна генерировать структурно корректные и правдоподобные ковариационные матрицы.

Таким образом, в работе продемонстрировано, что учет внутренней геометрии данных является ключевым фактором для успешного прогнозирования сложных систем временных рядов.

В качестве направлений для дальнейших исследований можно выделить:

- Исследование более сложных архитектур для score-сети и различных типов римановых многообразий.
- Разработка методов для оценки неопределенности прогноза путем анализа распределения, генерируемого стохастическим (SDE) семплером.

Программная реализация предложенных методов, скрипты для проведения экспериментов и генерации данных находятся в открытом доступе в репозитории на платформе GitHub¹.

Список литературы

- [1] Valentin De Bortoli, Émile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022.
- [2] James B Elsner and Anastasios A Tsonis. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenum Press, 1996.
- [3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [4] James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2012.
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.
- [6] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.
- [7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

¹<https://github.com/intsystems/2024-Project-152/tree/master/code>

- [8] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.
- [9] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- [10] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.
- [11] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. volume 25, pages 127–154, 2006.
- [12] P. Thomas Fletcher and Sarang Joshi. Riemannian geometry for the analysis of spherical images. *IEEE Transactions on Medical Imaging*, 26(7):955–962, 2007.
- [13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.