Правительство Российской Федерации Федеральное государственное автономное
образовательное учреждение высшего образования
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(государственный университет)

Кафедра ”Интеллектуальные системы”

Выпускная квалификационная работа магистра
На тему

Модели согласования скрытого пространства в задаче корреляционного анализа

Сафиуллина Роберта Руслановича

Научный руководитель
Стрижов В.В.
д.ф.-м.н.

Долгопрудный — 2023

Аннотация

Мы исследуем проблему прогнозирования одной траектории по другой. Мы должны найти зависимости в сигналах сложной структуры, используя линейные и нелинейные методы. Для этого предлагается уменьшить размерность исходной и целевой частных моделей, экономя пространство, и выполнить процедуру согласования модели в результирующем скрытом пространстве. В этой работе мы анализируем несколько методов многомодельного моделирования для построения функции согласования. Предлагаемое решение включает в себя выбор признаков с использованием подхода квадратичного программирования и метода частичных наименьших квадратов. Проверка результата будет осуществляться по сигналам от носимых устройств.

Содержание

1. Глава 1: Введение **4**

2. Глава 2: Обзор литературы **4**

3. Датасет **6**

4. Теоретические основы **7**

4.1 Основные понятия римановой геометрии 7

4.2 Динамические системы и их свойства 7

4.3 Корреляционный анализ в евклидовом пространстве 7

4.4 Переход к риманову пространству 8

5. Методология **8**

5.1 Обзор предлагаемого подхода 8

5.2 Предварительная обработка данных и их преобразование 8

5.3 Переход от евклидова пространства к риманову пространству 8

5.4 Корреляционный анализ в римановом пространстве 8

5.5 Метрики оценки и критерии эффективности 9

6. Вычислительный эксперимент **11**

Список используемой литературы **11**

1. Глава 1: Введение

Исследование представляет собой разработку новой методологии для определения корреляций и закономерностей в наборах данных временных рядов, используя уникальные свойства римановой геометрии. Это действительно актуально при анализе временных рядов, происходящих из разных динамических систем. В рамках данной работы, основной задачей является выявление зависимостей между траекториями и способностью прогнозировать одну траекторию на основе другой. Для достижения этой цели мы используем как линейные, так и нелинейные методы, включая снижение размерности исходного и целевого пространств с использованием частных моделей, а также согласование моделей в полученном скрытом пространстве.

В процессе исследования мы анализируем несколько методов множественного моделирования для построения функции согласования. В частности, предлагается использовать отбор признаков с помощью квадратичного программирования и метода частных наименьших квадратов. Для проверки валидности наших результатов мы используем данные, полученные от носимых устройств, что обеспечивает практическую применимость наших результатов.

В итоге, целью нашего исследования является разработка методологии корреляционного анализа траекторий и временных рядов, а также выявление наиболее эффективных методов множественного моделирования для построения функции согласования. Основными вопросами, которые мы хотим исследовать, являются выбор подходящих методов снижения размерности и оценка адекватности и точности выбранных методов множественного моделирования в условиях сложных структурных сигналов.

2. Глава 2: Обзор литературы

В данной главе представлен обзор существующих подходов к корреляционному анализу временных рядов, основанный на линейных и нелинейных методах. Корреляционный анализ использует коэффициент корреляции Пирсона (ρ) для измерения линейной связи между двумя переменными:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{1}$$

где $\text{cov}(X, Y)$ обозначает ковариацию между переменными X и Y, а σ_X и σ_Y обозначают их стандартные отклонения.

В контексте исследования сложных структурных сигналов актуально применение геометрии Римана и методов машинного обучения. Риманово многообразие определяется как дифференцируемое многообразие, снабженное римановой метрикой, которая задает инвариантное скалярное произведение на каждой точке многообразия:

$$g_p(u_p, v_p) = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}(p) u_p^i v_p^j, \tag{2}$$

где $g_p(u_p, v_p)$ обозначает риманову метрику, $g_{ij}(p)$ обозначает элементы матрицы метрики, а u_p^i и v_p^j обозначают координаты векторов u_p и v_p в точке p.

Основываясь на римановых многообразиях и нейронных сетях, разработаны различные подходы к снижению размерности и согласованию моделей. Приведены примеры применения этих методов в решении конкретных задач корреляционного анализа.

Также анализированы методы отбора признаков, включая квадратичное программирование и метод частных наименьших квадратов (PLS). Квадратичное программирование решает задачу оптимизации с квадратичной целевой функцией и линейными ограничениями:

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \quad \text{s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \tag{3}$$

где $\mathbf{x}$ является вектором переменных, $\mathbf{Q}$ - симметричной положительно определенной матрицей, $\mathbf{c}$ - вектором коэффициентов линейной части, а $\mathbf{A}$ и $\mathbf{b}$ - матрицей и вектором линейных ограничений соответственно.

Метод частных наименьших квадратов (PLS) используется для построения линейной регрессионной модели, основываясь на линейных комбинациях исходных переменных. Основная идея PLS заключается в построении новых переменных (латентных факторов) как линейных комбинаций исходных переменных, таким образом, чтобы максимизировать ковариацию между зависимой переменной и латентными факторами.

В частности, рассмотрена их эффективность и применимость в контексте корреляционного анализа сложных структурных сигналов, а также возможные альтернативные подходы.

В заключение главы формулированы основные выводы и определены вопросы, требующие дальнейшего изучения, а также заданы направления для разработки методологии корреляционного анализа на основе римановых пространств и машинного обучения.

Автокодировщики представляют собой мощный инструмент для снижения размерности данных и извлечения признаков. В данном параграфе рассматривается применение автокодировщиков, состоящих из полносвязных слоев, в контексте корреляционного анализа сложных структурных сигналов.

Пусть $\mathbf{x} \in \mathbb{R}^n$ - вектор признаков, полученный из временного ряда. Кодировщик преобразует этот вектор в сжатое представление $\mathbf{h} \in \mathbb{R}^m$, где $m < n$. Допустим, кодировщик имеет функцию активации ϕ и матрицу весов $\mathbf{W} \in \mathbb{R}^{m \times n}$. Тогда сжатое представление можно выразить следующим образом:

$$\mathbf{h} = \phi(\mathbf{W}x + b) \tag{4}$$

где $\mathbf{b} \in \mathbb{R}^n$ - вектор смещения.

Декодировщик, в свою очередь, пытается восстановить исходный вектор признаков из сжатого представления. Если у декодировщика есть функция активации ψ и матрица весов $\mathbf{V} \in \mathbb{R}^{n \times m}$, восстановленный вектор признаков $\hat{\mathbf{x}} \in \mathbb{R}^n$ может быть определен следующим образом:

$$\hat{\mathbf{x}} = \psi(\mathbf{V}x + c) \tag{5}$$

где $\mathbf{c} \in \mathbb{R}^n$ - вектор смещения.

Цель обучения автокодировщика состоит в минимизации функции потерь $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$, которая измеряет разницу между исходным вектором признаков и его восстановленной версией. Наиболее распространенным вариантом функции потерь является среднеквадратичная ошибка (MSE):

$$\text{MSE}(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2.$$

(6)

Применение полносвязных автокодировщиков в корреляционном анализе сложных структурных сигналов может привести к более эффективному и точному снижению размерности данных, что, в свою очередь, может упростить выявление зависимостей между различными временными рядами. Также сжатые представления, полученные с использованием автокодировщиков, могут быть использованы в качестве входных данных для других методов машинного обучения, таких как римановы многообразия и нейронные сети.

Одним из потенциальных направлений улучшения применимости автокодировщиков в данном контексте является использование различных архитектур, таких как сверточные автокодировщики или вариационные автокодировщики, которые могут привести к более эффективному сжатию и извлечению признаков, особенно для сигналов с пространственной или временной структурой.

В целом, применение полносвязных автокодировщиков в корреляционном анализе сложных структурных сигналов представляет собой перспективный подход для снижения размерности данных и выявления зависимостей между различными временными рядами. Исследование свойств и применимости автокодировщиков на полносвязных слоях может способствовать разработке новых методов и алгоритмов для анализа сложных структурных сигналов.

3. Датасет

WISDM Smartphone and Smartwatch Activity and Biometrics Dataset является обширным набором данных, включающим записи акселерометра и гироскопа с смартфонов и смарт-часов. Данный датасет предоставляет информацию о разнообразных физических активностях, выполняемых пользователями устройств, включая ходьбу, бег и другие. Это позволяет применить наш анализ временных рядов для изучения динамической системы, представленной человеком, и его взаимодействия с устройствами.

В рамках нашего исследования данный датасет может быть использован для сравнения и анализа временных рядов, полученных с акселерометров смартфонов и смарт-часов. Из-за разницы в размещении устройств на теле пользователя, эти два типа акселерометров предоставляют разные представления одной и той же динамической системы. Таким образом, можно исследовать взаимосвязь между данными с разных устройств, снижать размерность исходных данных и определить общие закономерности.

Применение корреляционного анализа временных рядов на основе Риманова пространства и машинного обучения к данным из WISDM Smartphone and Smartwatch Activity and Biometrics Dataset позволит нам выявить возможные зависимости между данными акселерометров разных устройств и оценить применимость нашего подхода в реальных условиях. Это также может привести к улучшению точности прогнозирования физической активности на основе данных акселерометра и повышению эффективности обнаружения аномалий или распознавания активностей.

4. Теоретические основы

4.1 Основные понятия римановой геометрии

Риманова геометрия является разделом дифференциальной геометрии, изучающим римановы многообразия и их свойства. Риманово многообразие определяется как дифференцируемое многообразие, на котором задана риманова метрика g, являющаяся симметричным (2,0)-тензором, который определяет скалярное произведение векторов касательного пространства.

Для двух векторов X и Y из касательного пространства в точке p на римановом многообразии M, риманова метрика определяется следующим образом:

$$g_p(X,Y) = \langle X, Y \rangle_p. \tag{7}$$

4.2 Динамические системы и их свойства

Динамическая система представляет собой математическую модель, описывающую эволюцию состояний системы во времени. Она состоит из фазового пространства X, элементы которого представляют состояния системы, и динамического закона ϕ, задающего эволюцию системы. Фазовое пространство может быть конечномерным или бесконечномерным, дискретным или непрерывным.

Для непрерывных динамических систем, представленных дифференциальными уравнениями, динамический закон имеет вид:

$$\frac{d}{dt}x(t) = F(x(t)), \tag{8}$$

где $F : X \to X$ - вектор-функция, определяющая динамику системы, $x(t) \in X$ - состояние системы в момент времени t.

4.3 Корреляционный анализ в евклидовом пространстве

Корреляционный анализ - это метод исследования статистической связи между двумя или более переменными. В контексте временных рядов, корреляционный анализ используется для измерения силы связи между двумя временными рядами.

Один из основных методов корреляционного анализа в евклидовом пространстве - это

коэффициент корреляции Пирсона, который определяется как:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{9}$$

где x_i и y_i - значения временных рядов, $\bar{x}$ и $\bar{y}$ - их средние значения, n - количество точек данных.

4.4 Переход к риманову пространству

Когда рассматриваются сложные структурные сигналы, такие как временные ряды из разных фазовых пространств, возникает необходимость перехода к риманову пространству. Римановы пространства позволяют учитывать геометрическую структуру данных, что может привести к улучшению результатов корреляционного анализа.

Для перехода к риманову пространству можно использовать различные методы, такие как построение римановых многообразий на основе евклидовых данных или использование специальных трансформаций.

5. Методология

5.1 Обзор предлагаемого подхода

В данной работе предлагается подход к корреляционному анализу временных рядов с использованием римановых пространств и машинного обучения. Основная идея состоит в переходе от евклидова пространства к риманову пространству, что позволяет учитывать геометрическую структуру данных и повышает эффективность корреляционного анализа.

5.2 Предварительная обработка данных и их преобразование

Перед анализом временных рядов производится предварительная обработка данных, включая удаление выбросов, нормализацию и сглаживание. Затем данные трансформируются с использованием методов снижения размерности, таких как автокодировщики, для получения компактного представления данных в пространстве меньшей размерности.

5.3 Переход от евклидова пространства к риманову пространству

Для перехода к риманову пространству используются различные методы, такие как построение римановых многообразий на основе евклидовых данных или применение специальных трансформаций. Это позволяет учитывать геометрическую структуру данных и повышает эффективность корреляционного анализа.

5.4 Корреляционный анализ в римановом пространстве

После перехода к риманову пространству, корреляционный анализ проводится с использованием геодезических расстояний и аффинных инвариантов, что позволяет учитывать геометрическую структуру данных и определить степень сходства между временными рядами. Можно использовать такие меры, как риманова корреляция, для измерения степени связи между временными рядами.

5.5 Метрики оценки и критерии эффективности

Для оценки эффективности предлагаемого подхода используются различные метрики и критерии, такие как коэффициент детерминации R^2, средняя абсолютная ошибка (MAE), средняя квадратическая ошибка (MSE) и другие. Эти метрики позволяют оценить степень сходства между исходными и предсказанными значениями временных рядов, а также определить точность и надежность предлагаемого подхода.

Таким образом, применение римановых пространств и машинного обучения для корреляционного анализа временных рядов позволяет учитывать геометрическую структуру данных и повышает эффективность анализа, что делает предлагаемый подход актуальным для исследования сложных структурных сигналов и прогнозирования динамических систем.

Теорема: Существование изометрической функции перехода между фазовыми пространствами одной динамической системы на римановых многообразиях

Пусть $\mathbf{M}_1$ и $\mathbf{M}_2$ - два римановых многообразия, представляющие фазовые пространства динамической системы, и пусть $\epsilon > 0$ - заданная точность аппроксимации. Мы предполагаем, что существует изометрическая функция перехода $\mathbf{T} : \mathbf{M}_1 \to \mathbf{M}_2$, такая что функция сохраняет риманову метрику и может быть аппроксимирована нейронной сетью с точностью ϵ, достигая взаимно однозначного преобразования между фазовыми пространствами.
Доказательство:

Определим непрерывную изометрическую функцию перехода $\mathbf{T} : \mathbf{M}_1 \to \mathbf{M}_2$ таким образом, чтобы для любых двух точек $\mathbf{x}, \mathbf{y} \in \mathbf{M}1$, риманово расстояние $\mathbf{dM}_1(\mathbf{x}, \mathbf{y})$ в $\mathbf{M}1$ было равно риманову расстоянию $\mathbf{dM}_2(\mathbf{T}(\mathbf{x}), \mathbf{T}(\mathbf{y}))$ в $\mathbf{M}_2$. Другими словами, мы требуем, чтобы $\mathbf{T}$ сохраняла риманову метрику.

Используя теорему об универсальном приближении для нейронных сетей, которая утверждает, что прямой нейронной сети с одним скрытым слоем, содержащим конечное количество нейронов, может аппроксимировать любую непрерывную функцию на компактных подмножествах $\mathbb{R}^n$ с любой желаемой точностью. Поскольку $\mathbf{T}$ является непрерывной функцией между римановыми многообразиями, мы можем построить нейронную сеть $\mathbf{f}$, которая ее аппроксимирует.

Пусть $\epsilon > 0$ - заданная точность аппроксимации. Тогда существует нейронная сеть $\mathbf{f}$ такая, что для всех $\mathbf{x} \in \mathbf{M}_1$ выполняется неравенство:

$$||T(x) - f(x)|| < \epsilon$$

Минимизируя функцию потерь

$$\mathbf{L}(\mathbf{f}) = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{M}1} |\mathbf{dM}1(\mathbf{x}, \mathbf{y}) - \mathbf{dM}_2(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))|^2$$

нейронная сеть $\mathbf{f}$ сходится к изометрической функции перехода $\mathbf{T}$, сохраняющей риманову метрику между фазовыми пространствами динамической системы.

6. Вычислительный эксперимент

Список литературы

[1] Dingari NC, Horowitz GL, Kang JW, Dasari RR, Barman I (2012) Raman Spectroscopy Provides a Powerful Diagnostic Tool for Accurate Determination of Albumin Glycation. PLoS ONE 7(2): e32406. https://doi.org/10.1371/journal.pone.0032406

[2] Rondeau P, Bourdon E (2011) The glycation of albumin: Structural and functional impacts. Biochimie 93: 645–658.

[3] Guthrow CE, Morris MA, Day JF, Thorpe SR, Baynes JW (1979) Enhanced nonenzymatic glucosylation of human serum albumin in diabetes mellitus. Proc Natl Acad Sci 76: 4258–4261.

[4] Pilot, R.; Signorini, R.; Durante, C.; Orian, L.; Bhamidipati, M.; Fabris, L. A Review on Surface-Enhanced Raman Scattering. Biosensors 2019, 9, 57. https://doi.org/10.3390/bios9020057

[5] Kosecki SM, Rodgers PT, Adams MB (2005) Glycemic monitoring in diabetes with sickle cell plus beta-thalassemia hemoglobinopathy. Ann Pharmacother 39: 1557–1560.

[6] Smith, C.; Karunaratne, S.; Badenhorst, P.; Cogan, N.; Spangenberg, G.; Smith, K. Machine Learning Algorithms to Predict Forage Nutritive Value of In Situ Perennial Ryegrass Plants Using Hyperspectral Canopy Reflectance Data. Remote Sens. 2020, 12, 928. https://doi.org/10.3390/rs12060928

[7] Natalia L. Nechaeva, Irina A. Boginskaya, Andrey V. Ivanov, Andrey K. Sarychev, Arkadiy V. Eremenko, Ilya A. Ryzhikov, Andrey N. Lagarkov, Ilya N. Kurochkin, Multiscale flaked silver SERS-substrate for glycated human albumin biosensing, Analytica Chimica Acta, Volume 1100, 2020, Pages 250-257, ISSN 0003-2670, https://doi.org/10.1016/j.aca.2019.11.072.

[8] Åsmund Rinnan, Frans van den Berg, Søren Balling Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry, Volume 28, Issue 10, 2009, Pages 1201-1222, ISSN 0165-9936, https://doi.org/10.1016/j.trac.2009.07.007.

[9] Вероятность и математическая статистика: Энциклопедия / Под ред. Ю.В.Прохорова. — М.: Большая российская энциклопедия, 2003. — 912 с.

[10] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[11] Li, T., Zhu, S. Ogihara, M. Using discriminant analysis for multi-class classification: an experimental investigation. Knowl Inf Syst 10, 453–472 (2006). https://doi.org/10.1007/s10115-006-0013-y

[12] Ryzhikova, E.; Ralbovsky, N.M.; Halámková, L.; Celmins, D.; Malone, P.; Molho, E.; Quinn, J.; Zimmerman, E.A.; Lednev, I.K. Multivariate Statistical Analysis of Surface Enhanced Raman Spectra of Human Serum for Alzheimer's Disease Diagnosis. Appl. Sci. 2019, 9, 3256. https://doi.org/10.3390/app9163256

[13] Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010.

[14] M. Arslan, M. Guzel, M. Demirci and S. Ozdemir, ”SMOTE and Gaussian Noise Based Sensor Data Augmentation,”2019 4th International Conference on Computer Science and Engineering (UBMK), 2019, pp. 1-5, doi: 10.1109/UBMK.2019.8907003.

[15] Boginskaya, I.; Sedova, M.; Baburin, A.; Afanas’ev, K.; Zverev, A.; Echeistov, V.; Ryzhkov, V.; Rodionov, I.; Tonanaiskii, B.; Ryzhikov, I.; Lagarkov, A. SERS-Active Substrates Nanoengineering Based on e-Beam Evaporated Self-Assembled Silver Films. Appl. Sci. 2019, 9, 3988. https://doi.org/10.3390/app9193988

[16] Nechaeva, Natalia Eremenko, Arkadiy Kurochkin, Ilya Boginskaya, Irina Afanasiev, Konstantin Ryzhikov, Ilya Sedova, Marina. (2019). Glicated human albumin registration using nanostructured silver substrates films realizing the effect of surface enhanced scatteringG. Mendeleev. 4. 10.32743/2658-6495.2019.4.4.206.