«Московский физико-технический институт (национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Курдюкова Антонина Дмитриевна

Снижение размерности фазового пространства в задачах канонического корреляционного анализа

03.03.01 – Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:

д.ф.-м.н. Стрижов Вадим Викторович

Москва

2022

Содержание

Аннотация

Данная работа посвящена задаче снижения размерности фазового пространства методами канонического корреляционного анализа. Исследуется связь между методом канонического корреляционного анализа и методом сходящихся перекрестных отображений Сугихары. Вид прогностических моделей представим в виде условия принадлежности двух аттракторов, восстанавливаемых в исходном и целевом фазовых пространствах, к общей динамической системе. В работе рассмотрены методы PLS-CCA, нелинейный метод CCA, seq2seq, Neural ODE. Сформулирован вариант теоремы о вложениях Такенса для проверки того, что метод канонического корреляционного анализа или другой метод прогноза удовлетворяет условиям Сугихары. Решается прикладная задача в теоретической постановке. Рассматривается видеоряд ходьбы человека с акселерометром на руке.

Ключевые слова: *снижение размерности, фазовое пространство, аттрактор, CCM, теорема Такенса о вложениях*

1 Введение

Решается задача прогнозирования сигналов походки человека. Такие сигнлалы обладают сложной структурой, то есть имеются зависимости и изменяющийся период. Рассматриваются два связанных пространства. Например, акселерометр и гироскоп одного мобильного устройства, два акселерометра в правой и левой руке человека, траектория движения руки, восстановленная по видеоряду движения человека, и акселеорметр на этой руке.

Для улучшения качества прогноза, а также для упрощения прогностической модели предлагается учесть зависимость между временными рядами, а также перейти в пространство меньшей размерности. Снижение размерности позволит учитывать внутреннее низкоразмерное проедставление временных рядов в прогностической модели.

Для определения наличия связи между временными рядами используется метод сходящегося перекрестного отображения (convergent cross mapping, CCM) [1, 2]. Метод CCM проверяет, насколько близки точки фазового пространства временного ряда s_1, соответствующие ближайшим соседям ряда s_2. Иными словами, существует ли взаимно однозначное соответствие, которое отображает окрестность фазовой траектории s_1 в окрестность фазовой траектории s_2.

Для снижения размерности траекторного пространства используются метод проекций на латентные структуры (partial least squares PLS) [3, 4], нелинейный PLS [5], seq2seq, NeuralODE. Снижение размерности позволяет сделать прогностическую модель более устойчивой, изучить связь между главными компонентами временных рядов, а также найти траекторное подпространство, в котором удастсться обнаружить связь медлу временными рядами.

В работе исследуется связь между методами корреляционного анализа и методом сходящегося перекрестного отображения. Для CCM нет способа выбора собственного подпространства, в котором аппроксимируется многообразие компакта и работает прогностическая модель. На текущий момент выбор собственного пространства осуществляется перебором по главным компонентам, например в [6]. Работа Исаченко [..] по PLS дает возможность перенести методы выбора подпространства с PLS на CCM.

Требуется построить прогностическую модель, связывающую метод сходящегогося перекрестного отображения и методы канонического корреляционного анализа. Целью работы является построение более простой моели, работающей не хуже уже существующих моделей прогнозирования временных рядов.

Определение 1 *Динамическая система – множество элементов, для которого задана функциональная зависимость между временем и положением в фазовом пространстве каждого элемента системы*

Динамическая система представляет собой такую математическую модель некоего объекта, процесса или явления, в которой пренебрегают «флуктуациями и всеми другими статистическими явлениям».

Определение 2 *Многообразие – хаусдорфово топологическое пространство со счётной базой, каждая точка которого обладает окрестностью, гомеоморфной евклидову пространству* $\mathbb{R}^n$

2 Постановка задачи

Дан временной ряд s_1. Значения временного ряда заданы в моменты времени $1, \ldots, N_1$. Требуется построить прогноз ряда на следующие m значений $N_1 + 1, \ldots, N_1 + m$.

При построении прогностической модели $\mathcal{F}$ нужно учесть влияние ряда $s_2 = \{s_i^2\}_{i=1}^{N_2}$ на ряд s_1. Значения ряда s_2 в моменты времени $N_1 + 1, \ldots, N_1 + m$ известны, то есть $N_2 > N_1 + m$.

Для построения прогноза ряда s_1 на один шаг по времени вперед будем учитывать L предыдущих значений этого ряда и все предшествующие текущему моменту времени значения ряда s_2. Тогда прогностическая модель имеет вид:

$$\widehat{s}\,^1_{t+1} = \mathcal{F}(\widehat{\mathbf{w}}, s_t^1, \ldots, s_{t-L+1}^1, s_1^2, \ldots, s_t^2), \tag{1}$$

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} S(\mathbf{w}, s_1, \widehat{s_1}),$$

где S — функция потерь.

3 Теоретическая часть

Пусть $s_1 = \{s_i^1\}_{i=1}^{N_1}$ и $s_2 = \{s_i^2\}_{i=1}^{N_2}$ — заданные временные ряды. Опишем, как строится фазовое пространство $\mathbf{X}$ временного ряда. Строится ганкелева матрица для ряда s_1:

$$\mathbf{H_1} = \begin{bmatrix} s_1 & \ldots & s_{n_1} \\ s_2 & \ldots & s_{n_1+1} \\ \vdots & \ddots & \vdots \\ s_{k_1} & \ldots & s_{N_1} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \mathbf{s}_1^1, \mathbf{s}_2^1, \ldots, \mathbf{s}_k^1 \end{bmatrix}, \quad k_1 = N_1 - n_1 + 1,$$

где n – ширина окна. Аналогично для временного ряда s_2. Тогда вектора $\mathbf{s}_1^1, \mathbf{s}_2^1, \ldots, \mathbf{s}_k^1$ образуют фазовую траекторию, или, иными словами, аттрактор $\mathbf{M}_1$ временного ряда s_1. На эти же вектора натянуто фазовое пространство $\mathbf{X_1}$ размерности n_1 временного ряда s_1. Формально, под терминами *фазовое пространство* и *аттрактор* будем понимать следующее:

Определение 3 *Фазовое пространство* $\mathbf{X}$ *динамической системы – совокупность всех допустимых состояний динамической системы.*

Определение 4 *Аттрактор* $\mathbf{M}$ *– компактное подмножество фазового пространства динамической системы, все траектории из некоторой окрестности которого стремятся к нему при времени, стремящемся к бесконечности.*

3.1 Метод сходящихся перекрестных отображений

Метод сходящихся перекрестных отображений (convergent cross mapping, CMM) используется для исследования временных рядов на нанличие причинно–следственной связи. Корелляция не подразумевает причинно–следственную связь между рядами. Метод основан на теореме Такенса о вложениях. В общем случае многообразие аттрактора динамической системы может быть восстановлено по одной наблюдаемой $\mathbf{X}$.

Согласно методу временной ряд s_1 может быть восстановлен по ряду s_2 только если временной ряд s_2 связан с рядом s_1. Временные ряды считаются связанными, если окрестность фазовой траектории $\mathbf{x}$ временного ряда s_1 взаимно однозначно отображается в окрестность фазовой траектории $\mathbf{y}$ ряда s_2. Иными словами,

Определение 5 *Аттракторы* $\mathbf{M}_1$ *и* $\mathbf{M}_2$ *наблюдаемых* $\mathbf{X}$ *и* $\mathbf{Y}$*, если* $\mathbf{X}$ *и* $\mathbf{Y}$ *принадлежат одной динамической системе.*

3.2 Метод проекций на латентные структуры

Метод проекций на латентные структуры PLS [3, 4] используют для нахождения фундаментальных зависимостей между двумя матрицами $\mathbf{X}$ и $\mathbf{Y}$. Отбираются наиболее значимые прихнаки. Новые признаки являются их линейными комбинациями. Осуществляется переход в фазовое пространство меньшей размерности. Метод PLS позволяет найти фазовое подпространство, в котором наблюдается связь между главными компонентами исходных временных рядов. Это позволяет исследовать наличие связи между временными рядами.

Пусть $\mathbf{X} \in \mathbb{R}^{m \times n}$ и $\mathbf{Y} \in \mathbb{R}^{m \times r}$ — матрицы двух фазовых пространств, построенных по временному ряду $\mathbf{s}_1$ и $\mathbf{s}_2$ соответственно. Требуется построить прогноз временного ряда $\mathbf{s}_2$ с учетом связи с временным рядом $\mathbf{s}_1$. Предполагается линейная зависимость между строками $\mathbf{X}$ и $\mathbf{Y}$:

$$\mathbf{Y}_i = \mathbf{X}_i \cdot \mathbf{\Theta} + \boldsymbol{\varepsilon} \quad \mathbf{Y}_i \in \mathbb{R}^r,\ \mathbf{Y}_i \in \mathbb{R}^n,\ i = 1, \dots, m, \tag{2}$$

где $\mathbf{\Theta}$ — матрица весов линейной зависимости, ε — вектор ошибок.

Ошибка вычисляется по формуле:

$$S(\mathbf{\Theta}, \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{X} \cdot \mathbf{\Theta}\|_2^2 \tag{3}$$

Алгоритм PLS находит матрицы $\mathbf{T}, \mathbf{P}, \mathbf{Q}$, с помощью которых осуществляется переход в латентное пространство согласно формулам:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathsf{T}} + \mathbf{F} \tag{4}$$

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^{\mathsf{T}} + \mathbf{E}$$

Матрица $\mathbf{T}$ наилучшим образом описывает $\mathbf{X}$ и $\mathbf{Y}$. Ее столбцы ортогональны. Матрицами $\mathbf{P}$ и $\mathbf{Q}$ определяется переход из латентного пространства в исходное. Матрицы $\mathbf{X}$ и $\mathbf{Y}$ — матрицы невязок.

Алгоритм PLS также позволяет определить матрицу $\mathbf{W}$, с помощью которой рассчитывается матрица весов $\mathbf{\Theta}$:

$$\mathbf{\Theta} = \mathbf{W}(\mathbf{P}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{Q}^{\mathsf{T}} \tag{5}$$

3.3 Вычислительный эксперимент

Вычислительный эксперимент проводился на данных [7]. Результаты эксперимента позволяют ответеть на вопрос: достаточно ли из двух сигналов, акселеорметра и гироскопа, какого-либо одного из них. Этого можно достичь в случае хорошего качества восстановления одного сигнала по второму. В данном эксперименте по сигналу акселеорметра восстанавливается сигнал гироскопа с помощью алгоритма PLS.

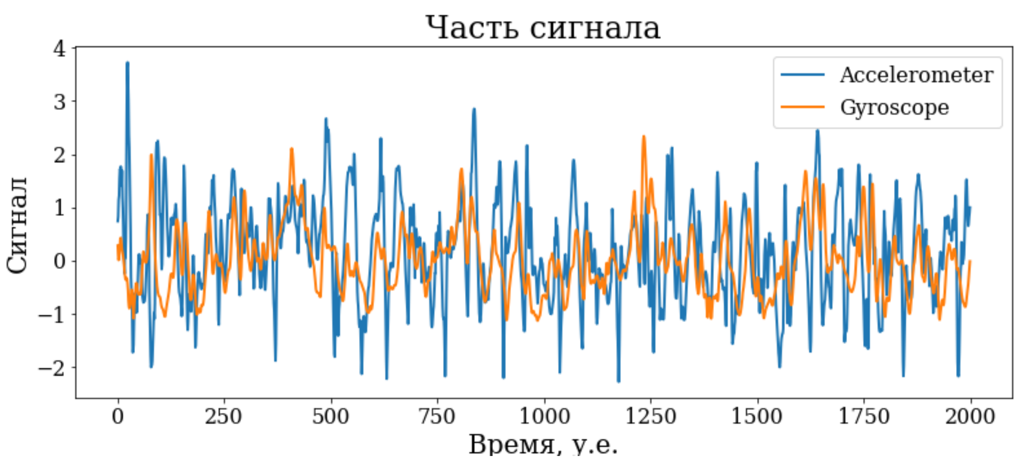


Рис. 1: Сигнал акселерометра и гироскопа походки человека

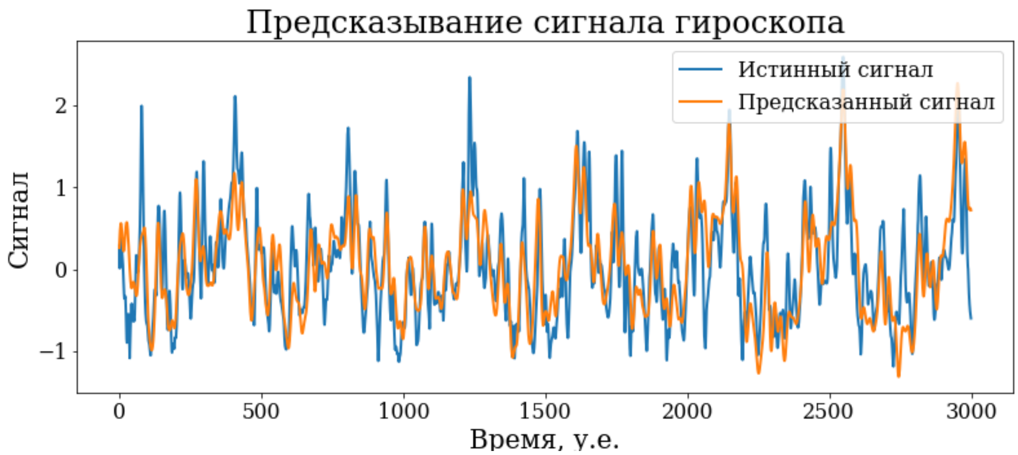


Рис. 2: Предсказание сигнала гироскопа по акселерометру, алгоритм PLS

4 Заключение

Список литературы

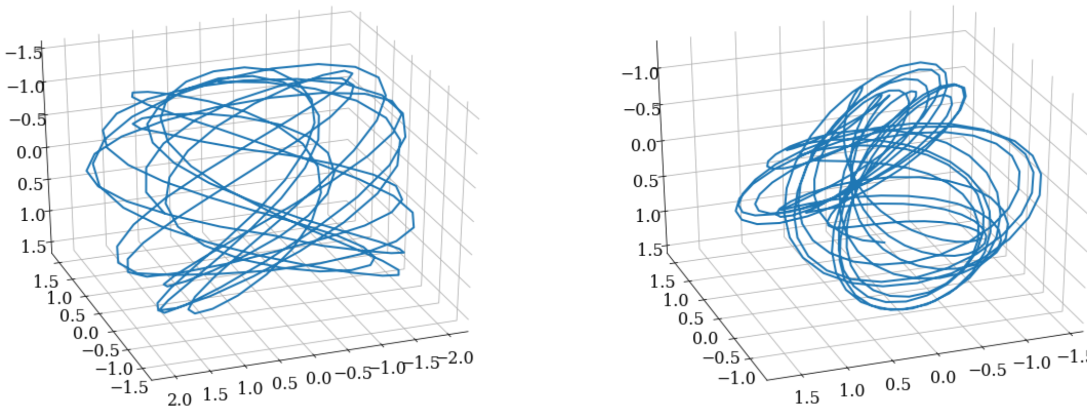
[1] George Sugihara and Robert M May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734–741, 1990.

[2] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

[3] Paul Geladi. Notes on the history and nature of partial least squares (pls) modelling. *Journal of Chemometrics*, 2(4):231–246, 1988.

[4] Agnar Höskuldsson. Pls regression methods. *Journal of chemometrics*, 2(3):211–228, 1988.

[5] Фарух Юрьевич Яушев, Роман Владимирович Исаченко, and Вадим Викторович Стрижов. Модели согласования скрытого пространства в задаче прогнозирования. *Системы и средства информатики*, 31(1):4–16, 2021.



(a) Акселерометр

(b) Гироскоп

Рис. 3: Фазовые траектории для первых трех главных компонент.

[6] Карина Равилевна Усманова and Вадим Викторович Стрижов. Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей. *Системы и средства информатики*, 29(2):12–30, 2019.

[7] Данные акселерометра и гироскопа: https://github.com/Intelligent-Systems-Phystech/Kurdyukova-BS-Thesis/tree/master/code/data.