Статистический анализ многомерных повторных наблюдений

Дамбаев Биликто Намсараевич, гр. 19.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к. ф.-м. н., доцент Алексеева Н.П.
Рецензент: младший научный сотрудник АО «Биокад»,
Мандрикова А.А.

Санкт-Петербург
2023г.

Введение

Саногенез — комплекс защитно-приспособительных механизмов, направленных на восстановление саморегуляции организма.

Рассматриваем модель **кривой саногенеза** (Барт, 2003), представляемую в виде двойной правой обратной функции $S(t) = e^{-\eta t} \cos \tau t$.

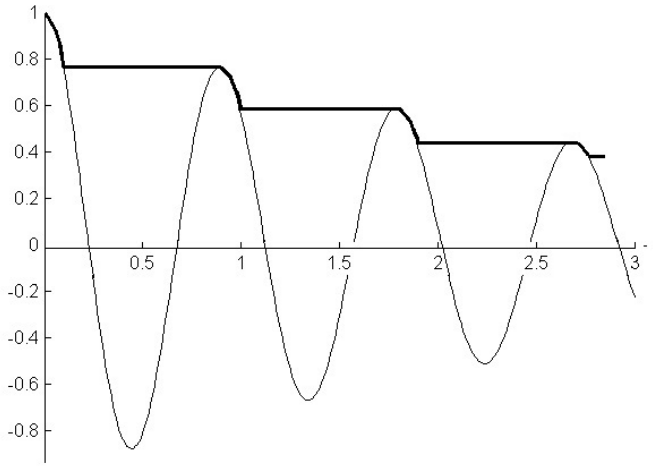

Рис.: Функция $S(t) = e^{-\eta t} \cos \tau t$ и ее двойная правая обратная.

Постановка задачи

$S(t) = e^{-\eta t} \cos \tau t$ — вещественная часть корреляционной функции КМНС[1] процесса.

Задачи:

- Выявить сочетания категориальных показателей, наибольшим образом влияющие на кривые саногенеза.
- Необходима статистика для проверки значимости отличия параметров кривых саногенеза.
- Выяснить, как отражается различие корреляционной структуры КМНС процесса на динамике линейной комбинации вещественной и мнимой его частей с наибольшей значимостью эффектов взаимодействия.

[1] Комплексный марковский нормальный стационарный процесс

Оценка параметров КМНС процесса

- $x_j(t) = u_j(t) + iv_j(t),\ j \in 1..n,\ t \in 1..k_j$, где k_j — число точек наблюдения у j-го индивида, n — число индивидов.
- $\mathbb{E}x_j(t) = 0$, ковариационная функция имеет вид

$$\mathcal{B}(t) = \mathbb{E}x_j(k)x_j(k+t)^* = \sigma^2 e^{-\eta|t| - i\tau t},\ \eta > 0.$$

- Обозначим через $r = e^{-\eta - i\tau}$, затем преобразуем параметр $\eta = -\ln\theta$ из $rr^* = \theta^2$.
- $A(l, m, t) = \sum_{j=1}^{n} \sum_{i=l}^{m} x_j(i)^* x_j(i+t)$

ОМП $\hat{\tau}, \hat{\eta}$ и их асимптотические дисперсии

Теорема (Алексеева, 2012)

Пусть $x_j(t) = u_j(t) + iv_j(t)$ *— n независимых реализаций КМНС процесса в* k_j *временных точках,* $N = \sum_{j=1}^{n} k_j$ *— общее число всех точек наблюдений,*

$$A_1 = A(1, k_j, 0),\ A_2 = A(2, k_j - 1, 0),\ A_3 = A(1, k_j - 1, 1),$$

$$\mathrm{Z}(\tau) = \mathrm{Re}(A_3)\cos\tau + \mathrm{Im}(A_3)\sin\tau.$$

Тогда для оценки $\hat{\tau}$ *выполняется*

$$\mathrm{tg}\,\hat{\tau} = \frac{\mathrm{Im}\,A_3}{\mathrm{Re}\,A_3},$$

а оценка $\hat{\theta} := e^{-\hat{\eta}}$ *является решением*

$$A_2(N-n)\theta^3 - \mathrm{Z}(\hat{\tau})(N-2n)\theta^2 - (NA_2 + nA_1)\theta + \mathrm{Z}(\hat{\tau})N = 0.$$

Асимптотические дисперсии оценок имеют вид

$$\mathbf{D}\hat{\tau} = \frac{1-\theta^2}{2\theta^2(N-n)},\ \mathbf{D}\hat{\theta} = \frac{N(1-\theta^2)}{2(N-n)(N-(N-2n)\theta^2)}.$$

Проверка однородности параметров

На основании полученных дисперсий и асимптотической нормальности ОМП получаем критерий для проверки однородности параметров.

Критерий проверки однородности параметров

Пусть $\zeta = \theta$ или τ. Тогда

$$H_0 : \zeta_1 = \zeta_2,$$
$$H_1 : \zeta_1 \neq \zeta_2,$$
$$z = \frac{\hat{\zeta}_1 - \hat{\zeta}_2}{\sqrt{\mathbf{D}\hat{\zeta}_1 + \mathbf{D}\hat{\zeta}_2}} \xrightarrow{D} \mathcal{N}(0,1).$$

Приложение метода к реальным данным

Решается задача систематизации многомерных повторяемых наблюдений при помощи сочетания модели кривых саногенеза и дисперсионного анализа на примере данных больных с алкогольно-абстинентным синдромом (ААС).

Используемые данные

Имеются показания разных метрик в течение реабилитационно-восстановительного периода алкоголиков.

- 34 индивида.
- 27 показателей — 18 категориальных и 9 количественных.
- 4 временные точки — 1, 2, 3 и 9 дни реабилитации.

Описание признаков. Выбор пар некоррелируемых признаков

Таблица: Описание основных признаков.

SBP	Артериальное систолическое давление
SV	Ударный объем сердца
Headache (H)	Наличие головной боли
Sweating (S)	Степень потоотделения
Tremor (T)	Степень тремора

Выбор **u** и **v**

Для построения оценок $\hat{\theta} = \hat{\theta}(\mathbf{u}, \mathbf{v})$ и $\hat{\tau} = \hat{\tau}(\mathbf{u}, \mathbf{v})$ выявлена пара некоррелируемых признаков на уровне значимости $\alpha = 0.1$:

$$(\mathbf{u}, \mathbf{v}) = (\text{SBP}, \text{SV})$$

Оценки параметров η, τ в зависимости от фактора H

Признаки	p-значение	Нет $(H = 0)$ $n = 21$	Есть $(H = 1)$ $n = 13$
(SBP,SV)	$H_0 : \theta_0 = \theta_1$ **0.024**	$\hat{\theta}_0 = 0.30 \pm 0.18$ $\hat{\tau}_0 = 0.11 \pm 0.59$	$\hat{\theta}_1 = 0.62 \pm 0.20$ $\hat{\tau}_1 = -0.28 \pm 0.30$

Таблица: Оценки $\hat{\theta} = e^{-\hat{\eta}}$ и $\hat{\tau}$ для $SBP + iSV$ по фактору H.

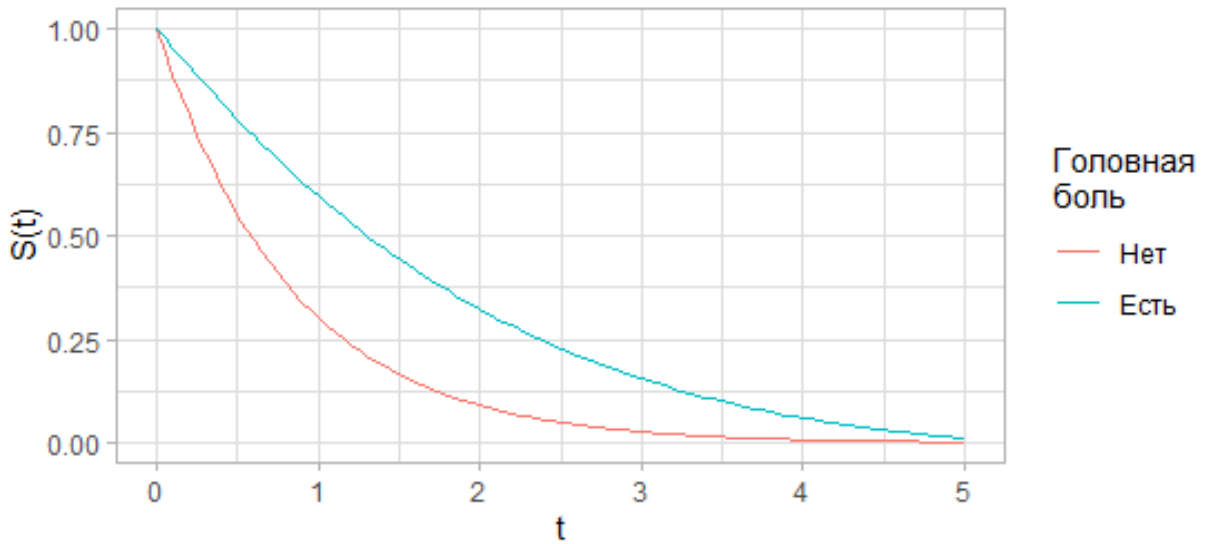

Рис.: График кривых саногенеза, построенных на паре SBP-SV и сгруппированных по фактору головной боли H

Задача поиска фактора наиболее значимого для различия кривых саногенеза

Из 18 категориальных переменных наибольшее различие по параметру θ получено по фактору H. Возникает вопрос — не будет ли более значимым какое-то сочетание факторов?

Одностороннее расстояние между кривыми саногенеза

Пусть $S^{(0)}$, $S^{(1)}$ — кривые саногенеза, построенные при условии $X = 0$ и $X = 1$ соответственно. Тогда

$$d(S^{(0)}, S^{(1)}) := (\theta_0 - \theta_1)^2$$

Симптомно-синдромальный метод

Симптом (Алексеева, 2021)

$\mathbb{X}_k = (X_0, X_1, \ldots, X_{k-1})$ — случайный k-мерный вектор с реализациями на $\{0, 1\}$.
Тогда симптомом называется $\mathcal{L}(\mathbb{X}_k) = \alpha_0 X_0 + \ldots + \alpha_{k-1} X_{k-1} \pmod 2$, где $\alpha_i \in \{0, 1\}$.

Вектор $(X_0, X_1, X_0 + X_1 \pmod 2)$ — синдром 1-го порядка.

Синдром $S(\mathbb{X}_{k+1})$ k-го порядка

Пусть $X_k \notin S(\mathbb{X}_k)$, а сложение определено покомпонентно
$S(\mathbb{X}_k) + X_k \pmod 2 := \{(S(\mathbb{X}_k))_i + X \mod 2\}_{i=1}^{2^k-1}$.
$S(\mathbb{X}_1) = S(X_0) = X_0$, $S(\mathbb{X}_2) = S(X_0, X_1) = (X_0, X_1, X_0 + X_1 \pmod 2)$. Тогда

$$S(\mathbb{X}_{k+1}) = (S(\mathbb{X}_k), X_k, S(\mathbb{X}_k) + X_k \pmod 2).$$

Симптомно-синдромальный метод

Синдром $V(\mathbb{X}_{k+1})$ k-го порядка

Пусть $X_k \notin V(\mathbb{X}_k)$, а умножение определено покомпонентно $V(\mathbb{X}_k)X_k := \{(V(\mathbb{X}_k))_i X\}_{i=1}^{2^k-1}$. $V(\mathbb{X}_1) = V(X_0) = X_0$, $V(\mathbb{X}_2) = V(X_0, X_1) = (X_0, X_1, X_0X_1)$. Тогда

$$V(\mathbb{X}_{k+1}) = (V(\mathbb{X}_k), X_k, V(\mathbb{X}_k)X_k).$$

Если в качестве базовых элементов $S(\mathbb{X}_k)$ использовать элементы $V(\mathbb{X}_k)$, то получим полиномы Жегалкина.

Суперсиндром $SV(\mathbb{X}_k)$

$SV(\mathbb{X}_k) = S(V(\mathbb{X}_k))$, размерности $2^K - 1$, где $K = 2^k - 1$

Применение симптомно-синдромальный метода

Наиболее значимый фактор

Пусть задан $\mathbb{X}$, $S^{(0)}$, $S^{(1)}$ — кривые саногенеза, построенные при условии $X = 0$ и $X = 1$ соответственно. Тогда X^*, такой что

$$X^* = \underset{X \in SV(\mathbb{X})}{\arg\max}\, d(S^{(0)}, S^{(1)}).$$

— наиболее значимый фактор.

$\mathbb{X} = (H, S, T)$, $\dim SV(\mathbb{X}) = 2^{2^3-1} - 1 = 127$.
Среди $SV(\mathbb{X})$ выделяем фактор $X^* = H + HT = H\overline{T}$, который характеризует церебральный/соматический вариант ААС[2].

[2] Алкогольно-абстинентный синдром

Оценки параметра θ в зависимости от фактора $H\overline{T}$

Признаки	p-значение	$H\overline{T} = 0$ $(n = 26)$	$H\overline{T} = 1$ $(n = 8)$
(SBP,SV)	$H_0 : \theta_0 = \theta_1$ **0.0003**	$\hat{\theta}_0 = 0.28 \pm 0.17$	$\hat{\theta}_1 = 0.78 \pm 0.22$

Таблица: Оценки $\hat{\theta} = e^{-\hat{\eta}}$ для $SBP + iSV$ при группировке по фактору $H\overline{T}$

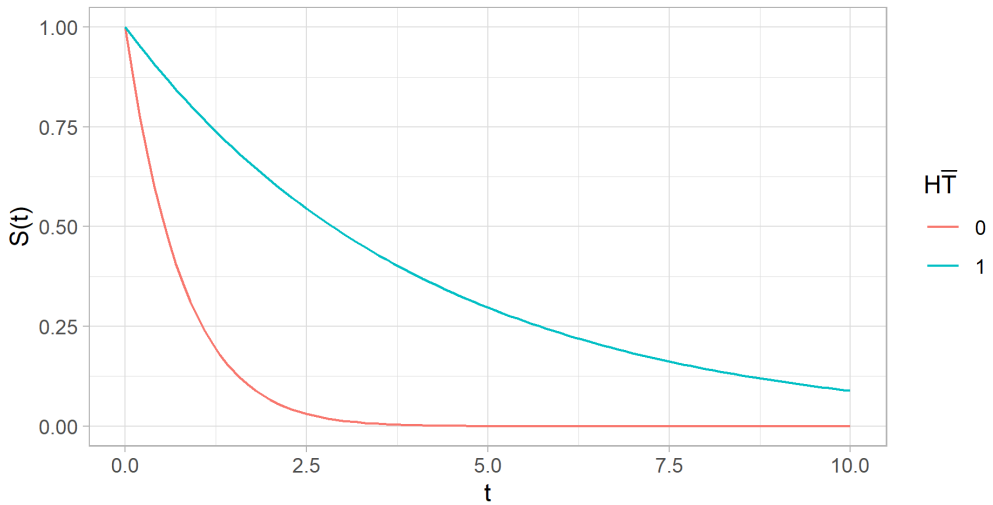

Рис.: График кривых саногенеза, построенных по паре SBP-SV и сгруппированных по фактору $H\overline{T}$.

Двухфакторная модель дисперсионного анализа с повторными наблюдениями

Модель

Пусть y_{ijt} наблюдение из i-ой группы, $i \in 1..I$, j-го индивида, $j \in 1..J$, в t-ый момент времени, $t \in 1..T$.

$$y_{ijt} = \mu + \alpha_i + e^1_{ij} + \beta_t + \gamma_{it} + e_{ijt}$$

- μ — генеральное среднее,
- α_i — фиксированный эффект группы,
- β_t — фиксированный эффект времени,
- γ_{it} — фиксированный эффект взаимодействия группы и времени,
- $e^1_{ij} \sim N(0, \sigma^2_1)$ — ошибка, связанная с разнообразием индивидов,
- $e_{ijt} \sim N(0, \sigma^2)$ — общая ошибка модели.

Многомерная дисперсионная модель

Модель:

$$Y_{rt} = y_{ijt}^{(r)} = \mu^{(r)} + \alpha_i^{(r)} + e_{ij}^{1(r)} + \beta_t^{(r)} + \gamma_{it}^{(r)} + e_{ijt}^{(r)}$$

$r \in 1..p$, где p — число моделей.

Задача: нахождение коэффициентов $a_1 \dots a_p$:

$$Z_t = \sum_{r=1}^{p} a_r Y_{rt} \quad t \in 1..T,$$

$$F_{AB}(a_1 \dots a_p) \to \max_{a_1 \dots a_p},$$

где F_{AB} — статистика критерия Фишера для проверки эффекта взаимодействия группы и времени.

Применение модели дисперсионного анализа

Согласно выбору пары ортогональных признаков, необходимо найти коэффициенты a_1, a_2 линейной комбинации

$$Z = a_1 Y_1 + a_2 Y_2,$$

где $(Y_1, Y_2) = (\mathrm{SBP}, \mathrm{SV})$, при которой различие во взаимодействии должно быть наибольшим.

Гипотеза:

$$H_0 : \gamma_{it} = 0 \quad \forall i \in 1, \ldots, \mathrm{I}, \ \forall t \in 1, \ldots, \mathrm{T}.$$

$$F_{AB}(a_1, a_2) \to \max_{a_1, a_2}.$$

Результаты MANOVA RM для $H\overline{T}$

Таблица: Минимальное p-значение эффекта взаимодействия $H\overline{T}$ и времени для пары SBP-SV

Признаки	p-значение	Коэффициенты линейной комбинации: (a_1, a_2)
SBP-SV	**0.015**	(-0.640,0.767)

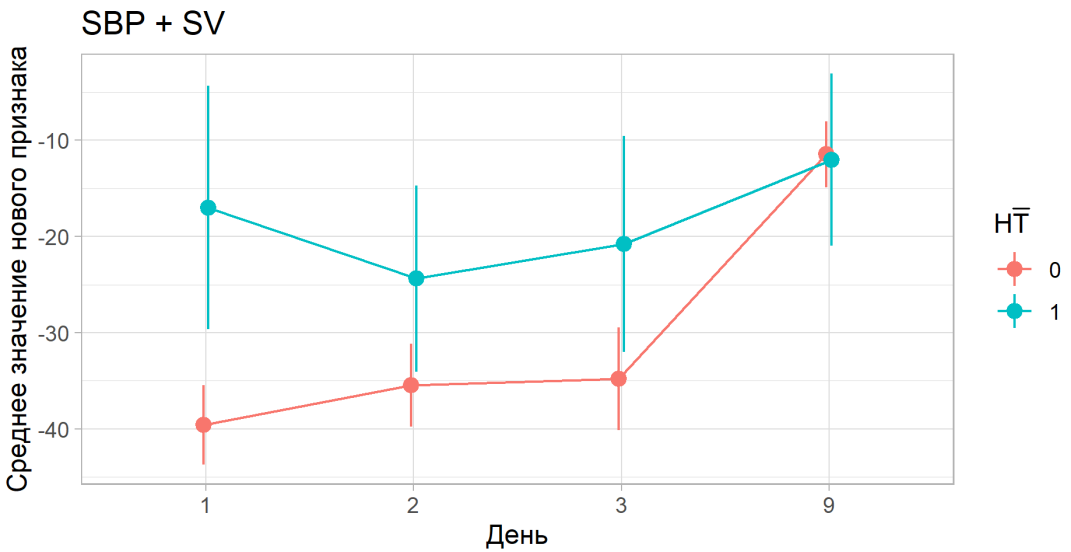

Рис.: Два типа динамики признака -0.64SBP$+0.767$SV (функционально стабильный и прогрессирующий).

Заключение

Результаты:

- Построен критерий для проверки однородности параметров кривых саногенеза.
- Разработан метод кластеризации данных по структуре корреляционных зависимостей повторных наблюдений на основе сочетания модели кривых саногенеза и дисперсионного анализа.
- В прикладном плане при помощи симптомного анализа выявлен латентный фактор $H\overline{T}$, характерный для разных форм алкогольно-абстинентного синдрома: церебрального и вегетативно-соматического варианта.