

# Статистический анализ многомерных лонгитюдных кардиологических данных

Орбидан Егор Владимирович  
науч. рук.: к.ф.-м.н., доцент Н.П. Алексеева  
рецензент: к.т.н., доцент Л.А. Белякова

Санкт-Петербургский Государственный Университет  
Кафедра статистического моделирования

2019

# Описание эксперимента [1]

Участники эксперимента — 9 студентов НГУ им. П.Ф.Лесгафта.

Показатели — данные мониторинга пяти гемодинамических характеристик в течение 3 минут в обычных условиях и 5 минут в условиях гипоксии.

Этапы исследования:

- 1 первоначальное перед трехмесячным курсом тренировок;
- 2 сразу после окончания тренировок;
- 3 через полгода.

Тренировки — чередование обычных условий с условиями гипоксии.

# Описание полученных экспериментатором признаков

- Основные:
  - SV — систолический объем;
  - EDP — конечное диастолическое давление;
  - SPR — удельное периферическое (артериальное) сопротивление сосудов;
  - HR — частота сердцебиения;
  - MBF — минутный объем крови (кровообращения);
- Специальные:
  - $MCI = tg(EDP/SV)$  — индекс, отражающий сократимость миокарда левого желудочка;
  - QRS — вектор деполяризации во фронтальной плоскости ЭКГ.

- Проверить, происходит ли адаптация к условиям гипоксии после продолжительных тренировок;
- Выделить группы индивидов по характеристикам, активизирующимся в экстремальных условиях;
- Выявить зависимость между основными характеристиками сердечно-сосудистой системы и новыми характеристиками, которые ввел экспериментатор;
- Упорядочить признаки по степени информативности в задаче классификации состояния испытуемого.

- Построить статистики Фишера для дисперсионного анализа с повторными измерениями со случайными эффектами для неполных данных;
- Выявить латентные характеристики по линейным комбинациям реализаций многомерных процессов, по которым достигается наиболее значимый эффект фактора взаимодействия;
- Построить кривые саногенеза для оценки параметров модели КМНС процесса;
- Статистические задачи: редуцировать данные, кластеризовать индивидов, упорядочить признаки по их информативности в классификации и выполнить канонический анализ.

# Дисперсионный анализ с повторными наблюдениями для неполных данных

$$x_{ijt} = \mu + \alpha_i + e_{ij}^1 + \beta_t + \gamma_{it} + e_{ijt},$$

- $\mu$  — генеральное среднее;
- $\alpha_i, \beta_t, \gamma_{it}$  — фиксированные эффекты группы, времени и взаимодействия,
- $e_{ij}^1 \sim N(0, \sigma_1^2), e_{ijt} \sim N(0, \sigma^2)$  — взаимно независимые ошибки.
- $I$  — число групп,  $T$  — число временных точек.
- $M_{it}$  — множество индивидов группы  $i$  во временной точке  $t$ ,  $m_{it}$  — его мощность.

# Матричная модель дисперсионного анализа для неполных данных

В работе [2] было показано, что

$y_{ijt} = x_{ijt} - x_{ij.} + H_{ij} + G_i = \beta_t + \gamma_{it} + \delta_{ijt}$ , где

- $H_{ij}$ ,  $G_i$  — индивидуальная и групповая поправки;
- $\delta_{ijt} = e_{ijt} - e_{ij.} + \varepsilon_{ij} + \epsilon_i$ ,  $\delta$  — вектор зависимых в совокупности ошибок с ковариационной матрицей  $\Lambda$ ;
- $\varepsilon_{ij}$  и  $\epsilon_i$  — ошибки, вызванные индивидуальной и групповой поправками соответственно.

В матричном виде модель можно записать как

$$Y = \mathbf{H}\Theta + \delta, \text{ где}$$

- $Y = (y_{111}, \dots, y_{1m_{11}1}, \dots, y_{I1T}, \dots, y_{Im_{IT}T})^T$  — вектор наблюдений;
- $\Theta = (\beta_1, \gamma_{11}, \dots, \gamma_{\iota 1}, \dots, \beta_\tau, \gamma_{1\tau}, \dots, \gamma_{\iota\tau})$  — вектор параметров размерности  $I(T-1)$ , где  $\tau = T-1, \iota = I-1$ .

Матрица  $\Lambda^{-1}$  содержит блоки  $\Lambda_{it,lk}^{-1}$  размерности  $m_{it} \times m_{lk}$ ,  $i, l = 1, \dots, I$ ;  $t, k = 1, \dots, T$ .

- $\lambda_{it,lk} = J_{it}^T \Lambda_{it,lk}^{-1} J_{lk}$ , где  $J_{it}$  — вектор из единиц размерности  $m_{it}$ ;
- $\lambda_{.t,.k} = \sum_{i=1}^I \sum_{l=1}^I \lambda_{it,lk}$ ;
- $GTr(\Lambda^{-1}) = \sum_{i=1}^I \sum_{t=1}^T \lambda_{it,it}$ ;
- $\mathbf{H}_b$  и  $\mathbf{H}_g$  — матрицы плана усеченных моделей;
- $R_0 = (Y - \mathbf{H}\hat{\Theta})^T \Lambda^{-1} (Y - \mathbf{H}\hat{\Theta})$ ,  
 $R_1 = (Y - \mathbf{H}_b \hat{\beta})^T \Lambda^{-1} (Y - \mathbf{H}_b \hat{\beta})$ ,  
 $R_2 = (Y - \mathbf{H}_g \hat{\gamma})^T \Lambda^{-1} (Y - \mathbf{H}_g \hat{\gamma})$ ;
- $\hat{\Theta} = (\mathbf{H}^T \Lambda^{-1} \mathbf{H})^{-1} \mathbf{H}^T \Lambda^{-1} Y$ ,  $\hat{\beta} = (\mathbf{H}_b^T \Lambda^{-1} \mathbf{H}_b)^{-1} \mathbf{H}_b^T \Lambda^{-1} Y$ .



# Математические ожидания $R_0, R_1, R_2$ для случайных эффектов

$$x_{ijt} = \mu + \alpha_i + e_{ij}^1 + b_t + g_{it} + e_{ijt}, \text{ где}$$

$b_t \sim N(0, \sigma_b^2)$ ,  $g_{it} \sim N(0, \sigma_g^2)$  — случайные эффекты времени и взаимодействия.

- $c_0 = N - I(T - 1)$ ,  $a_0 = (T - 1)(I - 1)$ ,  $b_0 = T - 1$ ;
- $a_1 = GTr(\Lambda^{-1} - \Lambda^{-1}\mathbf{H}_b(\mathbf{H}_b^T\Lambda^{-1}\mathbf{H}_b)^{-1}\mathbf{H}_b^T\Lambda^{-1})$ ;
- $b_1 = GTr(\Lambda^{-1}\mathbf{H}_b(\mathbf{H}_b^T\Lambda^{-1}\mathbf{H}_b)^{-1}\mathbf{H}_b^T\Lambda^{-1})$ ,  $b_2 = \sum_{t=1}^T \lambda_{..t}$ ;

## Утверждение

- 1  $\mathbb{E}(R_0) = c_0\sigma^2$ .
- 2  $\mathbb{E}(R_1 - R_0) = a_0\sigma^2 + a_1\sigma_g^2$ .
- 3  $\mathbb{E}(R_2 - R_0) = b_0\sigma^2 + b_1\sigma_g^2 + b_2\sigma_b^2$ .

# Проверка гипотез о незначимости факторов взаимодействия и времени

Гипотеза	Статистика	Распределение
$H_0 : \sigma_g^2 = 0$	$F_g = \frac{(R_1 - R_0)/a_0}{R_0/c_0}$	$F(a_0, c_0)$
$H_0 : \sigma_b^2 = 0$	$F_b = \frac{\frac{R_2 - R_0}{b_0} - d}{(R_1 - R_0)/a_0}, d = \left(\frac{b_1}{b_0} - \frac{a_1}{a_0}\right)$	$F(b_0, a_0)$

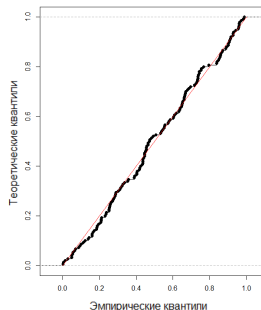
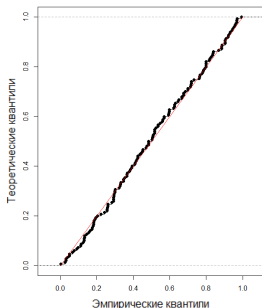


Рис.: Равномерность p-value для  $F_g$  (слева) и  $F_b$  (справа)

# Многомерная модель со случайными эффектами

## Модель:

$$x_{ijt}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + e_{ij}^{1(k)} + b_t^{(k)} + g_{it}^{(k)} + e_{ijt}^{(k)},$$

$k = 1, \dots, K$ , где  $K$  — число моделей.

**Задача:** поиск  $c_1, \dots, c_K$ ,  $c_k \in [0, 1]$ ,  $\sum_{k=1}^K c_k = 1$ :

$$\sum_{k=1}^K c_k x_{ijt}^{(k)} \rightarrow \max_{c_1, \dots, c_K} F_g$$

# Результат поиска латентных характеристик по реализациям многомерных процессов

Индивид	SV	EDP	HR	QRS
Dol	<b>0.7</b>	0.1	0.1	0.1
Sav	0.1	0.1	0.2	<b>0.6</b>
She	0.1	0.1	0.2	<b>0.6</b>
Sir	0.1	0.1	0.1	<b>0.7</b>
Hid	0.3	0.1	<b>0.5</b>	0.1
Ism	0.1	0.2	<b>0.6</b>	0.1
Rub	0.1	0.1	<b>0.7</b>	0.1
Shu	0.1	0.1	<b>0.7</b>	0.1
Spi	0.1	0.1	<b>0.6</b>	0.2

- SV — систолический объем;
- EDP — конечное диастолическое давление;
- HR — частота сердцебиения;
- QRS — вектор деполяризации во фронтальной плоскости ЭКГ.

# Оценка параметров корреляционной функции КМНС процесса [3, 4]

- $x_j(t) = \alpha_j(t) + i\beta_j(t)$  — некоррелированные признаки, где  $j = 1, \dots, n$ ,  $t = 1, \dots, k_j$ , где  $k_j$  — число первых точек наблюдения у  $j$ -го индивида;
- $S(t) = e^{-\eta t} \cos(\tau t)$  — вещественная часть корреляционной функции;
- оценка параметра  $\tau$ :

$$\operatorname{tg} \hat{\tau} = \frac{\operatorname{Im}(A_3)}{\operatorname{Re}(A_3)},$$

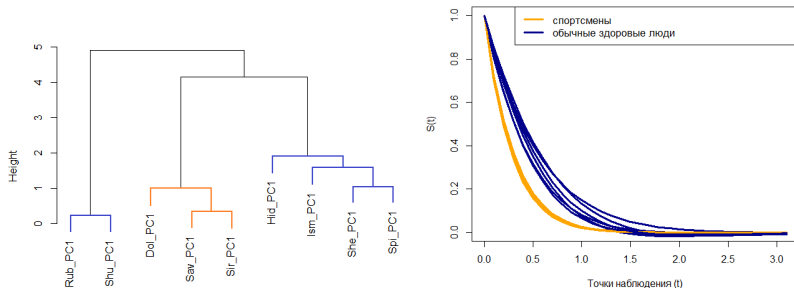
где  $A_3 = \sum_{j=1}^n \sum_{i=1}^{k_j-1} x_j^*(i) x_j(i+1)$ ;

- оценки остальных параметров имеют вид:

$$\hat{\eta} = -\ln \frac{Z(\hat{\tau})k}{A_1(k-1)}, \hat{\sigma}^2 = \frac{A_1}{kn},$$

где  $A_1 = \sum_{j=1}^n \sum_{i=1}^{k_j} x_j^*(i) x_j(i)$ ,  
 $Z(\tau) = \operatorname{Re}(A_3) \cos(\tau) + \operatorname{Im}(A_3) \sin(\tau)$ .

# Сравнение результатов кластеризации и кривых саногенеза



**Рис.:** Кластеризация и кривые саногенеза по данным за первые 2 этапа эксперимента (до и после тренировки)

# Результаты применения канонического корреляционного анализа

**Таблица:** Факторные нагрузки канонической величины специальных признаков (слева) и основных признаков (справа)

	$V_1$	$V_2$
MCI	<b>1</b>	<b>-0.03</b>
QRS	<b>-0.003</b>	<b>-1</b>

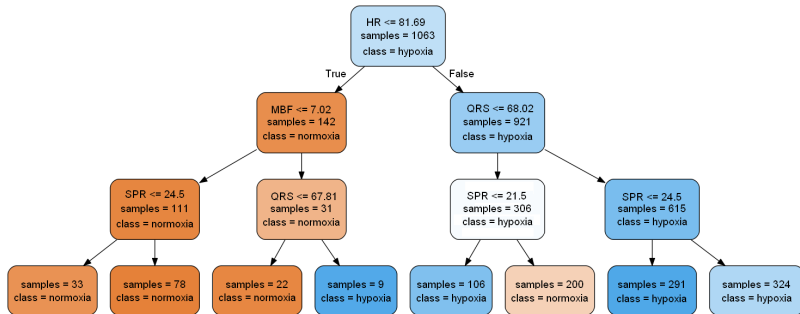
	$V_1$	$V_2$
SV	<b>-0.7</b>	<b>-0.05</b>
EDP	<b>0.7</b>	<b>-0.1</b>
SPR	<b>0.5</b>	<b>-0.06</b>
HR	<b>0.2</b>	<b>0.3</b>
MBF	<b>-0.5</b>	<b>0.06</b>

## Выводы:

- Переменная MCI сильно коррелирует с основными переменными, но при этом зависимости от QRS нет;
- Переменная QRS не зависит от основных переменных.

# Применение алгоритма случайного леса





- Классы: 0 — нормоксия, 1 — гипоксия;
- Критерий информативности — индекс Джини.



**Наиболее важные признаки:** **HR** (частота сердцебиения), **SPR** (удельное периферическое сопротивление сосудов), **QRS** (вектор деполяризации во фронтальной плоскости ЭКГ), **MBF** (минутный объем крови).



- Построены статистики для проверки гипотез в модели со случайными эффектами для неполных лонгитюдных данных;
- Разработан алгоритм оценки параметров линейной комбинации реализаций многомерного процесса с наиболее значимым эффектом фактора взаимодействия для выявления ведущих факторов, активирующихся в экстремальных условиях;
- Оценены параметры кривых саногенеза, ассоциированные с эффективностью тренировочного процесса;
- Подтверждена информационно-статистическая значимость специальных характеристик на основе канонического корреляционного анализа;
- Предложен метод упорядочивания признаков по их значимости для задачи классификации состояний при наличии и без гипоксии.

-  Радченко А. С. Взаимодействие пред- и постнагрузки сердца и RR интервалов при нормобарическом жестком гипоксическом воздействии у молодых здоровых лиц // Обзоры по клинической фармакологии и лекарственной терапии — 2013.
-  Alexeyeva N. Dual balance correction in repeated measures anova with missing data // Electronic Journal of Applied Statistical Analysis. — 2017.
-  Алексеева Н. П. Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия. — 2012.
-  Барт А. Г. Анализ медико-биологических систем. Метод частично обратных функций. — 2003.