Методы многомерного шкалирования в задачах прогнозирования клинических испытаний

Бородина Кристина Владимировна

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., профессор Ермаков М.С.
Рецензент: к.ф.-м.н., доцент Алексеева Н.П.

Санкт-Петербург
2015г.

Описание эксперимента. Постановка задачи

В исследовании участвовало 265 больных опийной наркоманией, которых разбили на три группы в зависимости от полученных препаратов:

1. PO/PI — плацебо орально и плацебо имплант;
2. NO/PI — налтрексон орально и плацебо имплант;
3. PO/NI — плацебо орально и имплант налтрексона.

Что нам известно о пациентах:

- пол;
- возраст;
- продолжительность лечения (в неделях);
- генотип.

Исследовательская задача

Всестороннее статистическое исследование влияния указанных факторов на эффективность лечения героиновой зависимости.

План исследования

1. Критерий согласия хи-квадрат для изучения зависимости от генов и их комбинаций второго ранга.
2. Логистическая регрессия для изучения зависимости от пола, возраста и типа терапии.
3. Решение задачи классификации:
   - методом многомерного шкалирования;
   - методом главных компонент.
4. Исследование влияния новых признаков на эффективность лечения в двумерном пространстве:
   - двухфакторным дисперсионным анализом для изучения зависимости новых признаков от типа терапии и исхода программы;
   - GEE/GLM подходом для изучения зависимости новых признаков от возраста, пола и времени лечения.
5. Симптомный анализ для изучения зависимости от различных комбинаций генов.

Критерий согласия хи-квадрат

Используется для выявления связи между категориальными переменными.

Нулевая гипотеза: ξ и η — независимы.
Статистика критерия:

$$t = \sum_{i=1}^{k} \sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n})^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}},$$

- n_{ij} — количество индивидов с i-ей градацией признака ξ и j-ой градацией признака η,
- $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$,
- $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$.

Таблица: Значимость влияния генов на выход из программы

	1	2	3	4	5	4⊕5	1⊕2	3⊕4	1⊕5
p-value	0,49	0,47	0,68	0,07	0,41	0,03	0,49	0,52	0,78

Логистическая регрессия

- Рассматривается множество независимых вещественных переменных $x_1, \ldots, x_n$, на основе значений которых требуется вычислить вероятность принятия значения 0 или 1 зависимой переменной y.

Вероятность наступления события $y = 1$ равна:

$$P(y = 1|x) = \pi(\beta^T x),$$

где $z = \beta^T x$, β — векторы-столбец коэффициентов регрессии.

$$\pi(z) = \frac{1}{1+e^z}.$$

Функцию распределения y при заданном x можно записать в таком виде:

$$P(y|x) = \pi(\beta^T x)^y (1 - \pi(\beta^T x))^{1-y}, y \in \{0, 1\}.$$

Таблица: Значимость влияния признаков на выход из программы

	group	age	gender
p-value	5,6e-09	0,917	0,076

Многомерное шкалирование

Преимущество метода

Позволяет расположить индивидов в метрическом пространстве выбранной нами размерности, получив в качестве исходных данных матрицу сходства объектов произвольного типа.

Цель

Изучить метод многомерного шкалирования с точки зрения других статистических методов.

- **На входе**: матрица с категориальными признаками размерности $n \times m$ (n—число индивидов, m—число признаков).
- Каждому признаку сопоставим бинарный индикатор $G_j, j = 1, \ldots, n$ размерности $n \times k_j$, где k_j—количество категорий каждого признака.
- **На выходе**:
  X — матрица оценок объектов размерности $n \times p$,
  Y_j — матрица категорий квантификаций размерности $k_j \times p$.

Функция потерь:

$$\sum_{j=1}^{m} tr(X - G_j Y_j)^T (X^{(t)} - G_j Y_j) \to \min_{\{X\},\{Y_j\}}.$$

Применение метода многомерного шкалирования

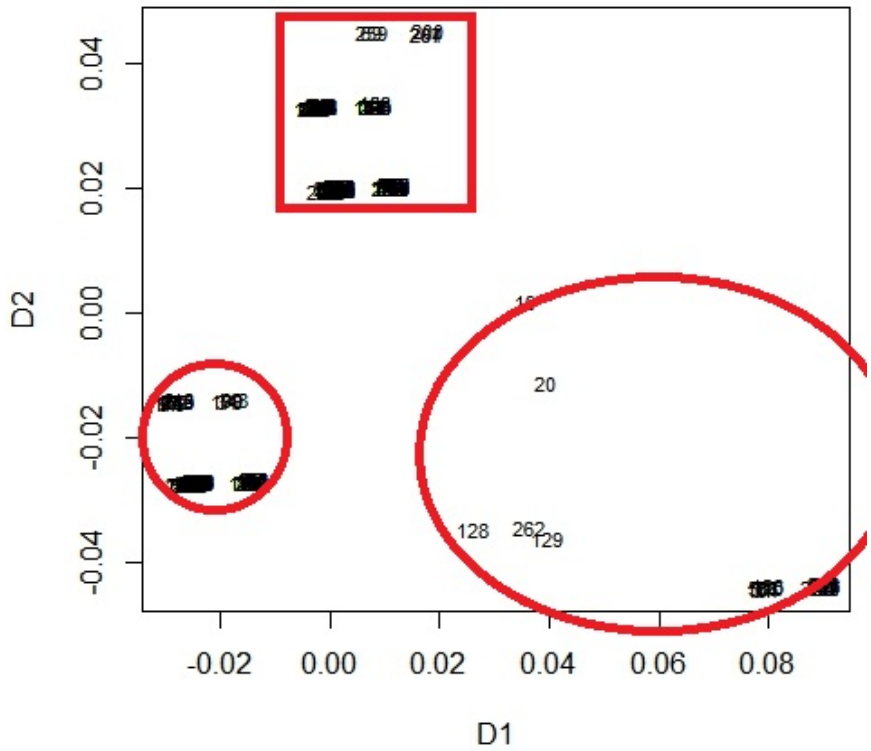

Таблица: Значимость влияния признаков на кластеры

	group	gender	endpore1	agegroup
p-value	0,529	0,967	0,883	0,685

Кластеризация на основе информационного разнообразия групп

- Пусть имеется группа n индивидов, характеризуемая признаком ξ со значениями $x_1, \ldots, x_m$.
- Пусть $n = a_1 + \cdots + a_m$, где a_i — количество индивидов со значением x_i признака ξ.

Информационное разнообразие I этой группы:

$$I = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

Информационный выигрыш от слияния двух групп A и B с разнообразиями I_A и I_B равен

$$\Delta I = I_{A+B} - I_A - I_B.$$

Таблица: Таблица сопряженности кластеров многомерного шкалирования и информационного разнообразия

	1	2	3
1	103	0	0
2	0	139	0
3	3	2	20

Метод главных компонент

Применяется для снижения размерности пространства наблюдаемых векторов, не приводя к существенной потере информативности.

На входе: числовые признаки $f_1(x), \ldots, f_n(x)$.
На выходе: новые признаки $g_1(x), \ldots, g_m(x)$, $(m < n)$.

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, j = 1, \ldots, n,$$

$$\sum_{i=1}^{l} \sum_{j=1}^{n} (\hat{f}_j(x_i) - f_j(x_i))^2 \to \min_{\{g_s(x_i)\},\{u_{js}\}},$$

где u_{js} — значения матрицы нагрузки, $x_1, \ldots, x_l$ — обучающая выборка.

Применение метода главных компонент

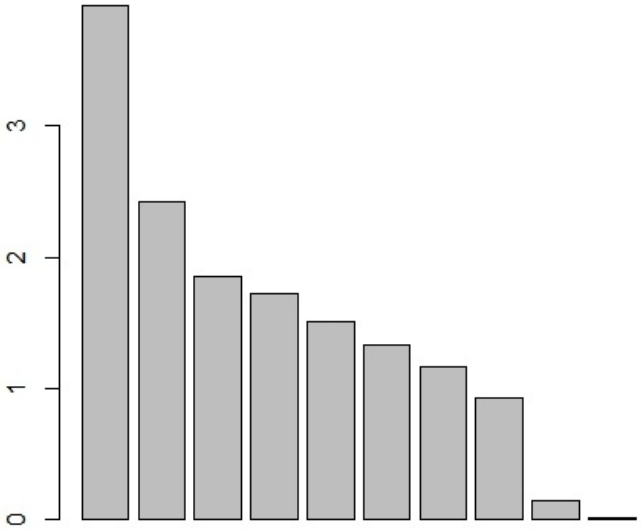

Рис.: *Вклад каждой компоненты в разброс данных*

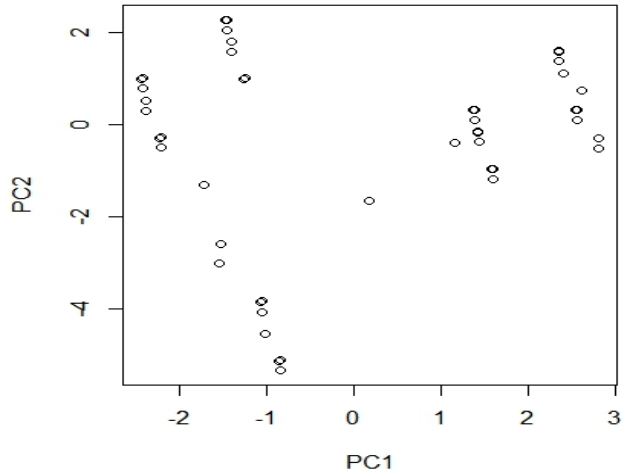

Рис.: *Расположение индивидов на графике первых двух компонент*

Таблица: Коэффициенты корреляции новых признаков

	D1	D2
PC1	0,32	-0,89
PC2	0,94	0,29

Двухфакторный дисперсионный анализ

Модель двухфакторного дисперсионного анализа:

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \xi_{ijk},$$

- A, B — факторы;
- x_{ijk} значение переменной X;
- μ — генеральное среднее;
- α_i — эффект фактора A;
- β_j — эффект фактора B;
- $(\alpha\beta)_{ij}$ — эффект взаимодействия факторов A, B;
- ξ_{ijk} — ошибки, независимые и нормально распределенные.

Таблица: Значения p-value в дисперсионном анализе

	D1	D2	PC1	PC2
group	0,502	0,387	0,264	0,537
endpore1	0,172	0,986	0,664	0,29
group*endpore1	0,411	0,798	0,625	0,47

GEE подход

Рассмотрим следующую задачу: на каждом из $i = 1, \cdots, N$ объекте сделано n_i измерений $y_i = (y_{i1}, \cdots, y_{in_i})$.

Пусть

$$\mu = E(y) = g(x^T \beta),$$

$$Var(y) = \phi v(\mu),$$

- x — матрица исходных данных (независимых признаков),
- β —p-мерный вектор коэффициентов регрессии,
- g — так называемая функция связи,
- $Var(y)$ — дисперсия y,
- ϕ - параметр масштаба,
- v — функция дисперсии.

Обобщенные уравнения оценок имеют вид:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \Sigma(y_i)^{-1} (y_i - \mu_i(\beta)) = 0,$$

где $\Sigma(y_i)$ — это матрица ковариации.

Применение GEE подхода

Таблица: Значения p-value в GEE/GLM подходе

	D1	D2	PC1	PC2
group	0.51	0,45	0,59	0,47
age	0,73	0,66	0,57	0,61
gender	0,45	0,19	0,26	0,10
endw	0,13	0,51	0,90	0,14

D1, D2 — новые признаки, полученные многомерным шкалированием;
PC1, PC2 — новые признаки, полученные методом главных компонент.

Таблица: Значения p-value в GEE/GLM подходе для индивидов с group=0

	D1	D2	PC1	PC2
gender	0,093	0,311	0,074	0,063
endpore1	0,403	0,005	0,392	0,149

Понятие симптома

- Случайный вектор $X = (X_1, \ldots, X_m)^T$ со значениями 0 или 1.
- Множество Ω_m размерности 2^m возможных значений вектора $x = (x_1, \ldots, x_m)^T, x_i \in F_2$.

Определение

Обозначим k–подмножество из m натуральных чисел через $\tau = (t_1, \ldots, t_k) \subseteq (1, 2, \ldots, m)$ и зададим вектор–строку $A_\tau = (a_1, \ldots, a_m)$ с компонентами

$$a_i = \begin{cases} 1, & i \in \tau \\ 0, & i \notin \tau \end{cases}$$

Линейная комбинация вида $X_\tau = A_\tau X (mod 2)$ называется **симптомом ранга k.**

Алгоритм отбора симптомов

Задача

Выявление совокупности генетических факторов, значимо влияющих на тяжесть наркотической зависимости.

- Выбор признаков индикатора рецессивности генотипов.
- Построение линейных комбинаций признаков над F_2 (симптомов) с ограничением на ранг $k < 4$.
- Исследование влияния симптомов на результат лечения в качестве фактора на всей выборке и на кластерах, полученных в результате многомерного шкалирования
  - в модели правого цензурирования;
  - в модели интервального цензурирования.
- Вычисление вероятности ошибки классификации.

Анализ результатов

- Правое цензурирование. Индикатор — результат выполнения программы.

Таблица: Значимость эффекта симптома ранга 3 в модели правого цензурирования

симптом	p-value(0)	p-value(2)	энтропия
$1 \oplus 4 \oplus 5$	0,0126	0,6618	0,2724
$3 \oplus 4 \oplus 5$	0,0246	0,8123	0,2686

- Интервальное цензурирование. Наблюдаемая величина $[t_1, t_2]$; t_1 —точка последнего наблюдения, $t_2 = t_1 + 1$.

Таблица: Значимость эффекта симптома ранга 3 в модели интервального цензурирования

симптом	p-value(0)	p-value(2)	энтропия
$1 \oplus 4 \oplus 5$	0,0186	0,7376	0,2724
$3 \oplus 4 \oplus 5$	0,0329	0,9860	0,2686

Анализ результатов

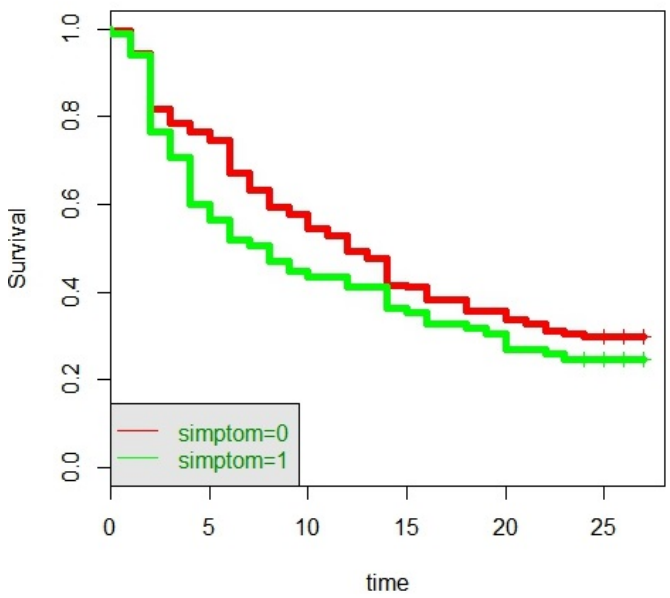

Рис.: *Симптом* $1 \oplus 4 \oplus 5$

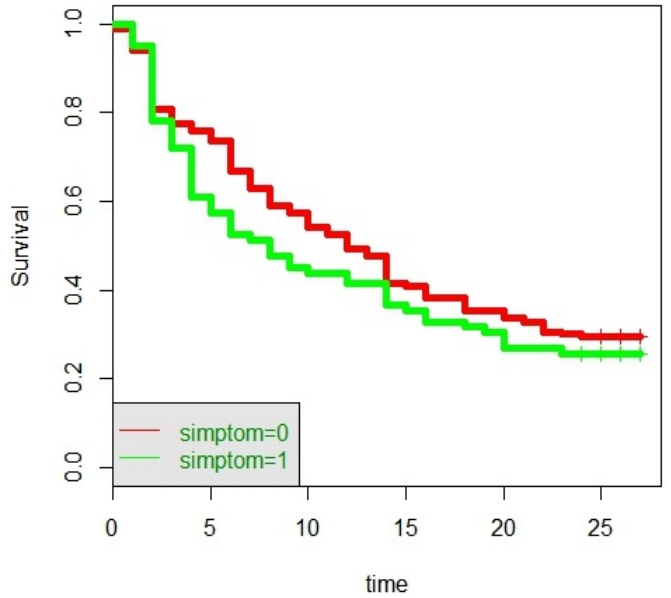

Рис.: *Симптом* $3 \oplus 4 \oplus 5$

- $1 \oplus 4 \oplus 5$—сочетание генов, отвечающее за никотиновую зависимость, когнитивные задачи и дисфорию;
- $3 \oplus 4 \oplus 5$—сочетание генов, отвечающее за импульсивность, когнитивные задачи и дисфорию.

Вывод

Все препараты снимают влияние генетических факторов.

Анализ результатов

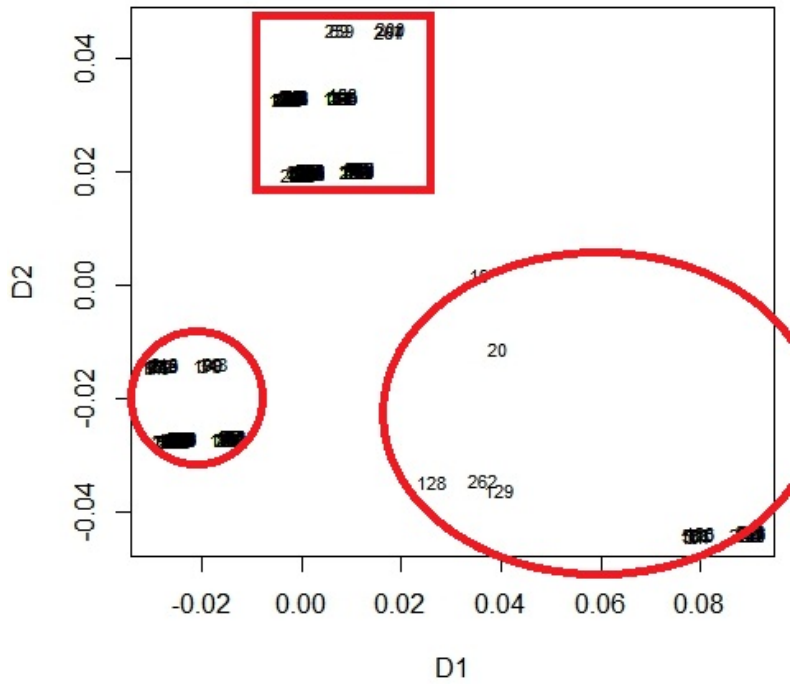

Таблица: Значимые симптомы ранга 3 многомерного шкалирования в модели правого (интервального) цензурирования

кластер 1	кластер 2	кластер 3
нет значимых	$1 \oplus 4 \oplus 5$	нет значимых
нет значимых	$3 \oplus 4 \oplus 5$	нет значимых

Вероятность ошибки классификации

- Разобьем всю выборку объемом $n = 265$ на тренировочную и тестовую.
- Вычислим вероятность ошибки классификатора на тестовой выборке

	$1 \oplus 4 \oplus 5$	$3 \oplus 4 \oplus 5$
на всей выборке	0,564	0,512
на 2-ом кластере	0,576	0,576

Вывод

Вероятность ошибки во всех случаях оказалась очень большой, что не дает нам возможности полагаться на полученные результаты классификации.

Результаты

В ходе данной работы было сделано следующее:

1. проведено первичное исследование исходных данных с помощью критерия хи-квадрат и логистической регрессии;
2. реализован алгоритм в языке программирования R, который позволил из исходных категориальных данных получить новые признаки $D1$, $D2$ и $PC1$, $PC2$;
3. исследованы новые признаки с помощью двухфакторного дисперсионного анализа и GLM подхода, последний реализован на языке R;
4. построен алгоритм для выявления линейных комбинаций над конечным полем факторов рецессивности генов (симптомов);
5. в перспективе проверить гипотезу, что аналогом многомерного шкалирования является симбиоз метода главных компонент и кластерного анализа на основе информационного разнообразия.